Flexible Spatial Models on the Example of Temperature in China

Master thesis submitted to

Prof. Dr. Ostap Okhrin

Prof. Dr. Wolfgang K. Härdle

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E. Centre for Applied Statistics and Economics

Humboldt-Universität zu Berlin

by

Anastasija Tetereva

(541990)

in partial fulfillment of the requirements
for the degree of

Master (M.Sc) in Statistics

Berlin, November 12, 2012

Acknowledgment

First of all, I would like to acknowledge the enthusiastic supervision of Prof. Dr. Ostap Okhrin. Without his guidance and persistent help this master thesis would not have been possible. It gives me great pleasure to express my appreciation to Prof. Dr. Wolfgang K. Härdle for the support and critical comments. I share the credit of my work with Prof. Dr. Brenda López Cabrera and the whole Ladislaus von Bortkiewicz Chair of Statistics.

The thesis would not have come to a successful completion, without the help I received from the staff of the *Collaborative Research Center 649: Economic Risk* and the *Research Data Center*.

My sincere thanks are extended to Prof. Dr. Jānis Valeinis for the encouragement and motivation.

I cannot find the words to express my gratitude to my parents, Larisa Puķe and Vasiliy Teterev, for their love, care and patience.

Deep gratitude is extended to my friends, Ilze, Elena, Elen and Tessa for being with me all the time. Their support and encouragement have been of great value for me.

Last but not least I appreciate the support of the DAAD, that made my studies in Berlin possible.

Abstract

Spatial modeling of temperature is of crucial importance for agriculture, industry and ecology. This work presents interpolation methods for the daily average temperature in China in the time period from 1957 to 2009. Due to complex topography and diverse climate of the country flexibility of the spatial models is of great importance. This study attempts to develop techniques which are able to minimize the spatial prediction error and to capture temperature extremes. The current research extends copula-based interpolation method and proposes the innovative IDW-GEV model. Spatial regression, kriging and inverse distance interpolation are used as a benchmark to evaluate the performance of suggested techniques.

Keywords: interpolation, kriging, inverse distance interpolation, copula, generalized extreme value distribution

Zusammenfassung

Die räumliche Modellierung der Temperatur ist von entscheidender Bedeutung für die Landwirtschaft, Industrie und Ökologie. In dieser Arbeit werden Interpolationsmethoden für die tägliche Durchschnittstemperatur in China in der Zeitspanne von 1957 bis 2009 präsentiert. Aufgrund der komplexen Topographie und unterschiedlichen Klimas des Landes ist die Flexibilität der räumlichen Modelle von großer Wichtigkeit. Diese Studie versucht Techniken zu entwickeln, die den räumlichen Vorhersagefehler minimieren und Temperaturextreme erfassen können. Die aktuelle Forschung erweitert die Kopula-basierte Interpolationsmethode und schlägt das innovative IDW-GEV Modell vor. Räumliche Regression, Kriging und inverse Distanzwichtung werden als ein Benchmark benutzt, um die Leistungsfähigkeit der vorgeschlagenen Techniken zu beurteilen.

Schlagwörter: Interpolationsverfahren, inversive Distanzwichtung, Kriging, Kopula, Extremwertverteilung

Contents

List of Figures

List of Tables

1 Introduction

Investigation of the weather phenomena has become increasingly popular in recent years. Average daily temperature is an input to the great amount of models in ecology, hydrology, geology and industry. It is of high importance to know the temperature at high spatial resolution to be able to predict plant productivity, snow melting process, crop responses to climate change, soil properties, heating and cooling energy demand. This task becomes especially challenging in the areas with vast landscape and diverse climate. However, the location of weather stations is often sparse and it is necessary to develop interpolation techniques which are able to predict the temperature in unknown locations with the smallest possible error. Moreover, these models should be flexible enough to be able to capture extreme observations.

The considerable amount of literature has been published on kriging, e.g. Holdaway (1992), Cressie (1991) and Diggle and Ribeiro (2007a). A large and growing body of papers has investigated the possibility to model temperature as function of some geographical characteristics. Chai et al. (2002) and Lauren et al. (2002) suggest to use linear regression on geographical coordinates. Another popular methodology is inverse distance weighting. Loecher (2011) takes use of IDW to interpolate particulates in Europe, Babak and Deutsch (2008) argues that the right choice of the parameters in such kind of model significantly reduces the interpolation error. Bardossy (2011), Pebesma et al. (2011), Kazianka and Pilz (2010), Gräler et al. (2010) demonstrate that copula approach can be applied to the spatial interpolation problem. They introduce bivariate spatial copula and argue that it should be used to model dependence of extreme events. The majority of these studies applied copula to uncorrelated data or ignored the serial dependence.

The purpose of this research is to extend the model used by Kazianka and Pilz (2010) by making it more flexible and applicable to the serial correlated temperature data. Other methods, such as regression, inverse distance weighting and kriging, are critically examined and used as a benchmark to evaluate the performance of the innovative models.

The master thesis has been divided into four parts. First of all, we give a description of the data set and proceed by a brief overview of the interpolation task in Chapter 3. Chapter 4 begins by laying out the theoretical dimensions of the current research in the field of spatial interpolation and proceeds with the description of the proposed models. In Chapter 5 the

models, proposed in Chapter 4, are applied to the temperature data of China. All models are compared by calculating out-of-sample prediction's error. Finally, the conclusions are presented.

The computations for Chapter 5 were performed with `R` version 2.15.1. Packages `copula` by Yan (2007), `CDVine` by Brechmann and Schepsmeier (2011), `geoR` by Diggle and Ribeiro (2007b), `gstat` by Pebesma (2006), `intamap` by Pebesma et al. (2011), `googlemap` by Loecher (2010) were used.

2 About the Data

This chapter presents the data set and is followed by the discussion of descriptive statistics. This step of the analysis is of high importance and helps to understand the data in detail.

China is the world's second-largest country by land area and the third by total area. China's landscape is vast and diverse, with forest steppes, deserts and subtropical forests being prevalent in the wetter south near Southeast Asia. The territory of China lies between latitudes 18°N and 54°N, and longitudes 73°E and 135°E. China's landscape vary significantly across its vast width. In some parts there are extensive and densely populated alluvial plains. On the edges grasslands predominate. China's climate is mainly dominated by dry seasons and wet monsoons, which lead to pronounced temperature differences between winter and summer. The climate in China differs from region to region because of the country's highly complex topography. All these geographical and climatology characteristics of the investigated area makes the analysis and modeling of the temperature data complicated. Finding an appropriate method to capture all regional features can be a challenge. Therefore, more flexible techniques should be developed.

The data was provided by the Climatic Data Center of the National Meteorological Information Center. The data gives information about the average daily temperature in China in the time period from the 1st January 1957 till the 31th December 2009. The average temperature of the day is calculated as the average of the maximum and minimum temperature. The accuracy of the measurement is 0.1°C. The records come from 159 land weather stations which cover almost all the provinces in China. There are no observations from the Tibet (Xizang) and Jilin provinces. Weather stations which are located in the Xinjiang, Hunan and Neimongol provinces are widely spaced.

Originally the data set contained missing and unrealistic values. These 147 observations were replaced by the average of the previous and the next day measurements on the station. Days which correspond to the 29th of February were excluded. Thus, the analyzed data massive contains $53 \cdot 365 = 19345$ records for each of the 159 weather stations.

To explore the given data by the means of descriptive statistics all stations were grouped in 5 clusters and one station from each cluster was selected for further numerical analysis.

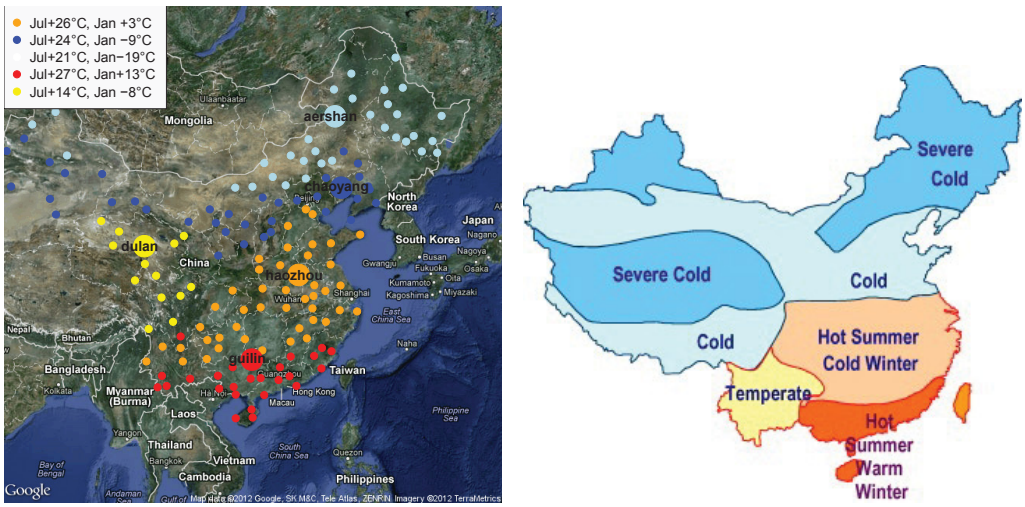

Figure 2.1: Weather stations in China grouped by clusters and climatic zones in China.
ClustGoog

Clusters were defined based on the average monthly temperature ($53 \cdot 12 = 636$ records for each station) using Ward algorithm. Distances between objects were calculated as Euclidean distances. The methods of cluster analysis are described in Härdle and Simar (2012). 5 cluster solution was chosen. The obtained clusters and the average temperature in January and July within each cluster can be seen in the figure 2.1. The same figure shows that the clusters agree with the climatic zones in China.

The picture 2.2 shows the temperature correlation matrix which uncovers high linear dependence among the stations. However, some stations have weaker correlation with the nearest stations than others. Thus, the interpolation task can be more challenging for some stations and precision of the results is expected to be region dependent.

Station	Min	Q1	Median	Mean	Q3	Max	SD
Chaoyang	-22.90	-2.50	11.10	9.13	21.00	33.40	12.90
Dulan	-21.10	-4.90	3.80	3.20	11.30	25.60	9.34
Aershan	-40.50	-16.70	-0.20	-2.58	11.90	27.60	15.66
Haozhou	-11.90	5.80	15.90	14.88	23.90	34.70	10.02
Guilin	-2.90	12.30	20.20	19.00	26.10	33.00	7.86

Table 2.1: Numerical summary for 5 weather stations.
NumSum

As was mentioned above, one station from each climatic zone (cluster) is chosen for the exploratory analysis. These stations are Aershan, Chaoyang, Dulan, Haozhou and Guilin.

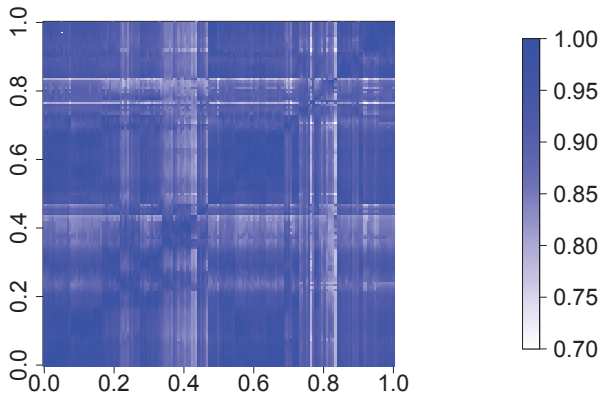

Figure 2.2: Correlation matrix for daily temperature.

CorMat

They correspond to big points in the figure 2.1. Table 2.1 gives basic numerical summary for 5 selected stations. In addition, boxplots by month are constructed and can be seen in the figure 2.3. Such a simple analysis makes possible to uncover regions with highest variance. It is expected that spatial interpolation in mountain areas will be less precise. It is worth to note that the variance is also month dependent and can cause higher interpolation error during the winter period. Therefore, prediction techniques should be adjusted taking into consideration that spatial stationarity is doubtful in this case. That is the reason for suggesting flexible local spatial interpolation methods and developing a number of new avenues for research.

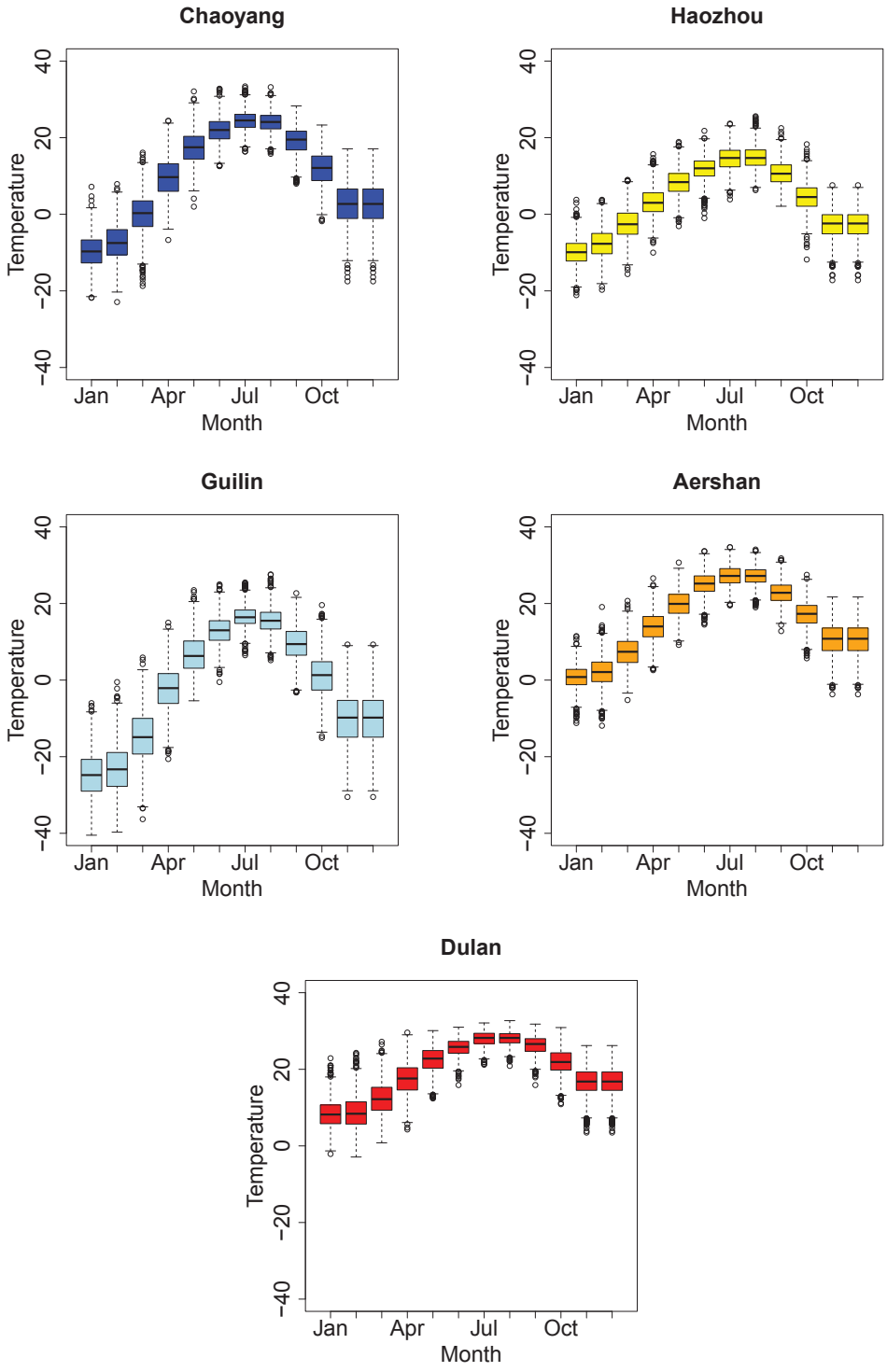

Figure 2.3: Boxplots of temperature in 5 weather stations grouped by month.

BoxPlots

3 Introduction to Spatial Interpolation

The current chapter gives a short introduction to spatial statistics and spatial interpolation problem. This field of statistics deals with the quantitative study of phenomena that are located in higher dimensional space. Spatial data analysis consists primarily of three main components: lattice data analysis, spatial point patterns modeling and geostatistics, whose objective is to create a continuous surface from a set of points. In other words, geostatistics is a part of spatial statistics which deals with data obtained by spatially discrete sampling of a spatially continuous process. More detailed information on spatial process is given in Diggle and Ribeiro (2007a).

The origins of geostatistics is found in mining type applications, however, nowadays it is widely used in modeling soil properties, ground water studies, rainfall precipitation, air pollution investigation, public health etc. For more details and many more examples we refer to Waller and Gotway (2004).

The interpolation task can be formalized as following: there are given n points in p dimensional subspace $S \subseteq \mathbb{R}^p$. Measurements of some process at these points at time t, $\{Z_t(x_i), i = 1, \ldots, n\}$, are available. The aim of the spatial interpolation is to describe spatial process $Z_t(x)$, $x \in S$ at each point of S and at each time moment t.

The main idea of spatial interpolation was formulated by Tobler (2008) - "Everything is related to everything else, but near things are more related than distant things". Thus, the prediction in unknown location is usually given by the weighted average of the nearest neighbors

$$\widehat{Z_t}(x_0) = \sum_{j=1}^{J} \lambda_j Z_t(x_j), t = 1, \ldots, T. \tag{3.1}$$

Here, $Z_t(x_j)$ are available measurements and λ_j are some weights, which vary according to interpolation technique. The number of neighbors and weights are calculated according to some criterion, e.g. minimizing 3.4, 3.5, 3.6 or 3.7. It is always possible to choose the number of neighbours or to fix the distance-based neighborhood. The illustration is given in the picture 3.1.

Spatial interpolation techniques can be divided into deterministic (e.g. inverse distance

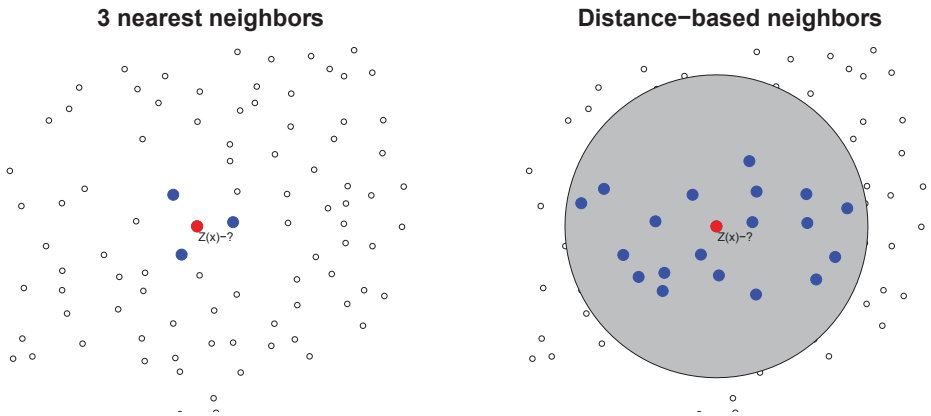

Figure 3.1: Illustration of nearest neighbor (left) and distance-based nearest neighbor (right) approach.

weighting, splines, radial basis functions etc.) and geostatistical (e.g. kriging, hierarchical models, copula etc.). The first one uses a mathematical function to calculate values in unknown locations and the obtained prediction is a deterministic number. The second method provides probabilistic estimates of the quality of the interpolation, e.g. variance. Interpolation can be exact and inexact. When inexact interpolation methods are used, prediction values can be different from the measured values.

Behavior of the spatial process $Z(x)$ can be completely characterized by its joint probability distribution. The type of stationarity plays a crucial role in spatial data analysis. The very strong kind of stationarity is one in which the joint probability function is invariant under some spatial shift h, i.e.

$$\begin{aligned} &\mathrm{P}\{Z(x_1) \leq z_1, Z(x_2) \leq z_2, \ldots, Z(x_p) \leq z_p\} \\ &= \mathrm{P}\{Z(x_1 + h) \leq z_1, Z(x_2 + h) \leq z_2, \ldots, Z(x_p + h) \leq z_p\}. \end{aligned}$$

Further on, for the simplicity of notation we will omit the time index t. Second order stationarity requires that the moments of the joint distribution do not change, i.e. $\mathrm{E}\{Z(x)\} = \mu, \forall x \in S$ and

$$\mathrm{Cov}\{Z(x), Z(x + h)\} = C(h), \forall x \in S. \tag{3.2}$$

$C(h)$ is called covariogram. The half of the variance of $Z(x_i) - Z(x_j)$ has a special name and is widely used characteristic of a spatial process. It is called semivariogram and is denoted $\gamma(x_i, x_j)$. Semivariogram multiplied by two is called variogram. In the case of stationary

process the variogram can be represented as a function of distance vector

$$\gamma(h) = \frac{1}{2}\,\mathrm{Var}\{Z(x) - Z(x+h)\}. \tag{3.3}$$

If the difference $Z_t(x) - Z_t(x+h)$ has constant mean and the variance depends only on h, instrinstic type of stationarity is met. If the process is furthermore isotropic, the variogram can be represented as a function of distance $\|h\|$. To simplify the notation further on we will write h instead of $\|h\|$, it is important to know that the absolute value of the distance is meant. Suggestions for further reading are Cressie (1991), Sherman (2011).

Another question which often arises in spatial interpolation is interpolation error. In order to be able to compare performance of competing spatial techniques and to find the optimal λ_j in 3.1 it is necessary to define the measure of prediction error. Some widely used measures are root mean squared error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{Z(x_i) - \widehat{Z}(x_i)\}^2}, \tag{3.4}$$

mean absolute error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Z(x_i) - \widehat{Z}(x_i)|, \tag{3.5}$$

mean absolute percentage error

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Z(x_i) - \widehat{Z}(x_i)|}{|Z_t(x_i)|}, \tag{3.6}$$

and mean error, which is useful to detect a bias

$$\mathrm{MA} = \frac{1}{n}\sum_{i=1}^{n}\{Z(x_i) - \widehat{Z}(x_i)\}. \tag{3.7}$$

Usually all the measures suggest to choose the same model. When the measure of error is chosen, crossvalidation technique should be used. The main idea of this procedure is to divide the sample into n folds, construct the model using $(n-1)$ folds and employ remaining 1 fold to calculate the error. The procedure is repeated n times. When n coincides with the sample size, the technique is called leave-one-out crossvalidation. More details on error measurement techniques can be found in Hastie et al. (2009).

4 Spatial Interpolation Models

4.1 Multiple Linear Regression

The objective of this section is simple, but effective, interpolation method - multiple linear regression (MLR). Details on MLR can be found in Green (2011). This method suggests to predict the value of the process in the unknown location as linear function of some geographical measures. MLR is extremely popular in meteorological application, e.g. in temperature or precipitation modeling. Such kind of research was done by Chai et al. (2002) and Lauren et al. (2002). The model is formulated as follows

$$Z(x_i) = \sum_{j=1}^{J} a_j \cdot g_j(x_i) + \varepsilon(x_i), i = 1, \ldots, n.$$

a_j are regression coefficients, $g_j(x_i)$ are geographical characteristics of x_i, e.g longitude, latitude, elevation, distance to the ocean, and $\varepsilon(x_i)$ is MLR residual. In some studies such fancy variables as slope of the mountain (e.g. north-facing side) are used. The advantage of the method is that extrapolation is possible (prediction of the phenomenon outside the observable region). However, assumptions of the MLR are rarely fulfilled in the context of spatial statistics. One more disadvantage is that sometimes only latitude and longitude of the location are available. We reger to Bivand et al. (2008) for further reading.

4.2 Inverse Distance Weighting

This section introduces another simple and intuitive spatial interpolation method - inverse distance weighting (IDW). The IDW prediction in unknown location $\widehat{Z}(x_0)$ is obtained as a weighted average of all available measurements. Usually weights are proportional to the inverse of the distance. Thus, closest available observations have stronger influence on prediction. The general formula of IDW is given by

$$\widehat{Z}(x_0) = \frac{\sum\limits_{j:\|x_j - x_0\| \leqslant h} w(x_j) Z(x_j)}{\sum\limits_{j:\|x_j - x_0\| \leqslant h} w(x_j)}, w(x_j) = 1/\|x_j - x_0\|^p, \tag{4.1}$$

where $w(x_j)$ are weights which are proportional to the distance between x_0 and x_j. Usually they are chosen as power function of Euclidean distance between two spatial points $\|x_j - x_0\|^{-p}$. Bivand et al. (2008) suggests to use $p = 2$.

However, this value is not always optimal. Babak and Deutsch (2008) show more statistical approach to the problem and choose p which minimizes RMSE given in 3.4 and prove by the means of empirical study that the interpolation error is strongly influenced by p.

Another parameter to be chosen is the number of neighbours or the distance for distance-based neighborhood (illustration is given in the chapter 3). It is intuitively clear that the weights for distant locations are very small, however, sometimes it makes sense to set them equal to zero and take into consideration only observations whose distance to x_0 does not exceed certain value. Thus, to obtain prediction with the smallest error, it is necessary to adjust simultaneously p and h.

The IDW interpolation is extremely simple and not computationally intensive. However, it has several pitfalls. IDW is deterministic interpolation method and does not provide variance of the prediction. In addition, standardized weights always lie between 0 and 1. Thus, interpolated values can not lie outside the range of observed values.

4.3 Kriging

This section explores some of the major method in geostatistics - kriging. The origins of this method are found in the year 1951, when a South African mining engineer Danie Gerhardus Krige published his thesis Krige (1951). At the beginning kriging was a weighted average of the data in the neighborhood of the point of interest. Since then kriging has been developed to tackle increasingly complex problems in many areas of geology, environmental science and public health. There are several kinds of kriging: simple, ordinary, universal, block, regression, cokriging etc. All of them are described in Cressie (1991). This research makes use of two most popular kriging procedures - ordinary and universal kriging. Ordinary kriging requires the second type stationarity, in particular, constant mean assumption

$$\mathrm{E}\{Z(x_i)\} = \mu, \ \forall \, x_i, i = 1, \ldots, n.$$

Universal kriging relaxes the assumption of the constant mean. It allows for spatial drift which is a deterministic function of some geographical characteristics of the location x_i, e.g.

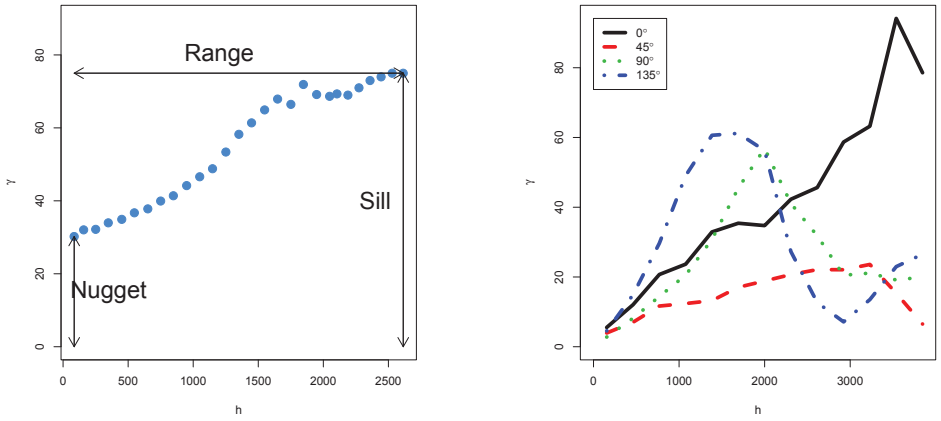

Figure 4.1: Illustration of isotropic and anisotropic variograms.

Illustr

longitude, altitude and distance to the ocean. This assumption can be formulated as

$$\mathrm{E}\{Z(x_i)\} = \sum_{j=1}^{J} a_j f_j(x_i), i = 1, \ldots, n.$$

Kriging procedure consists of two main parts. The first one is constructing the variogram 3.3. At the second step this knowledge is used to find the weights in the general interpolation formula 3.1. The most important object in geostatistics is the empirical variogram:

$$2\widehat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} \{Z(x) - Z(x+h)\}. \tag{4.2}$$

First look at the 4.2 can be misleading. If empirical variogram is calculated for each value of h, then the number of neighbors $N(h)$ is usually equal to 1 and the approximation of the variance by the empirical variance does not make sense. To avoid this problem, more precise formula should be used

$$2\widehat{\gamma}_{n(h)} = \frac{1}{\#N(h)} \sum_{(x_i,x_j)\in N(h)} \{Z(x_i) - Z(x_j)\}^2, h \in \mathbb{R}^p.$$

Here $N(h) = (x_i, x_j) : (r-\delta) \leq \|x_i - x_j\| \leq (r+\delta); i, j = 1, \ldots n, r = \|h\| > 0$. Definition is taken from Gaetan and Guyo (2010).

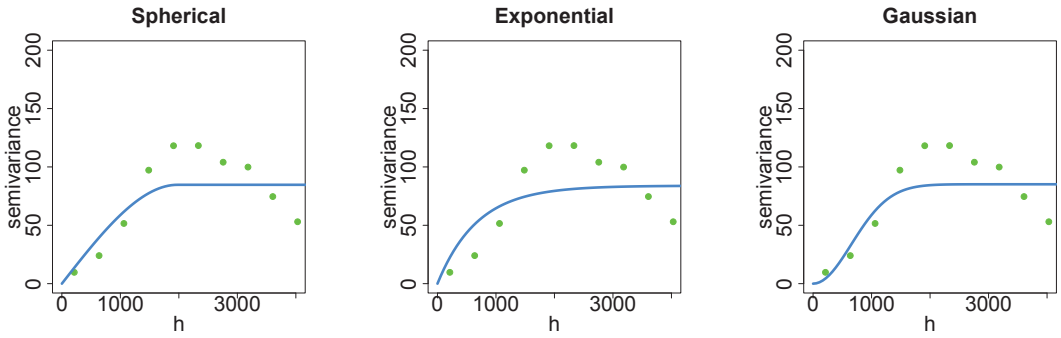

Figure 4.2: Three parametric variogram models.

VarType

Concepts of covariogram and variogram are closely related, i.e.

$$\gamma(h) = C(0) - C(h).$$

Proof is given in Gentona and Gorsich (2002). Definition of $C(h)$ is given in 3.2. It is intuitively clear that the covariance should decrease when the distance is increasing. Thus, it can be expected that the variogram is increasing function of the distance $\|h\|$. In the context of variogram, it is useful to introduce thee concepts: nugget, range and sill. Nugget is intercept of the variogram. It usually is referred to some unexplained variations and errors. Range is a scalar that controls the degree of correlation between data points, usually represented as a distance. Sill is value of the semivariance as the $\|h\|$ goes to infinity. It is equal to the total variance $C(0)$. It can be seen from 4.3 that $\gamma(h)$ reaches its maximal value $C(0)$ when $C(h) = 0$. Thus, range is the distance at which the variogram reaches the sill. It is the distance such that pairs of sites further than this distance apart are negligibly correlated. The empirical variogram and all above mentioned characteristics are depicted on the figure 4.1. Mathematical details on kriging theory can be found in Webster and Oliver (2007).

The next step after construction of the empirical variogram is to fit some parametric or nonparametric model. The most popular parametric models are Gaussian model

$$\gamma(h) = c + (s - c)\{1 - \exp\frac{-3h^2}{a^2}\}, \tag{4.4}$$

where c is nugget, s is sill and the range is given by a;
Spherical model

$$\gamma(h) = \begin{cases} c + (s - c)\{1.5\frac{h}{a} - 0.5\frac{h}{a}^3\} & \text{, if } h \leq a \\ c & \text{, otherwise} \end{cases}$$

exponential model

$$\gamma(h) = c + (s - c)\{1 - \exp\left(\frac{-3h}{a}\right)\}, \tag{4.5}$$

and linear model

$$\gamma(h) = c + bh^p. \tag{4.6}$$

All three functions are seen on the picture 4.2. The wide spread method for fitting the model to the empirical model is the least square optimization procedure which is described in Gaetan and Guyo (2010). Nonparametric methods are used as well, they are discussed in more details in Gentona and Gorsich (2012). The above discussion considers the case of isotropic data, i.e. sill and range values are the same, regardless of the direction being considered. This is not always the case. Empirical variograms for four directions are seen in the figure 4.1. Direction 45 degrees means that the dependence holds for direction 45 degrees with 22.5 degrees tolerance, i.e. angle between x_i and x_j lies between 22.5 and 67.5 degrees. When the empirical variogram is estimated and the theoretical model is fitted, it is possible to find the weights in the interpolation formula 3.1. The weights can be calculated according to

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} 0 & \widehat{\gamma}(x_1, x_2) & \cdots & \widehat{\gamma}(x_1, x_n) & 1 \\ \widehat{\gamma}(x_2, x_1) & 0 & \cdots & \widehat{\gamma}(x_2, x_n) & 1 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ \widehat{\gamma}(x_n, x_1) & 0 & \cdots & \widehat{\gamma}(x_n, x_n) & 1 \\ 1 & \cdots & \cdots & 1 & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} \widehat{\gamma}(x_1, x_0) \\ \widehat{\gamma}(x_2, x_0) \\ \vdots \\ \widehat{\gamma}(x_n, x_0) \\ 1 \end{bmatrix} \tag{4.7}$$

in the case of ordinary kriging. The interpolation formula 3.1 in matrix notation is given by the following equation

$$\widehat{Z}(x_0) = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix}^\top \times \begin{bmatrix} Z(x_1) \\ Z(x_2) \\ \vdots \\ Z(x_n) \end{bmatrix} \tag{4.8}$$

The universal kriging equation is given by

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} 0 & \cdots & \widehat{\gamma}(x_1, x_n) & 1 & f_1(x_1) & \cdots & f_m(x_1) \\ \widehat{\gamma}(x_2, x_1) & \cdots & \widehat{\gamma}(x_2, x_n) & 1 & f_1(x_2) & \cdots & f_m(x_2) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \vdots & 1 & 0 & 0 & \vdots & 0 \\ f_1(x_1) & \cdots & f_1(x_n) & 0 & 0 & \cdots & 0 \\ f_2(x_1) & \cdots & f_2(x_n) & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_m(x_1) & \cdots & f_m(x_n) & 0 & 0 & \cdots & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} \widehat{\gamma}(x_1, x_0) \\ \widehat{\gamma}(x_2, x_0) \\ \vdots \\ \widehat{\gamma}(x_n, x_0) \\ 1 \\ f_1(x_0) \\ \vdots \\ f_m(x_0) \end{bmatrix} \tag{4.9}$$

Detailed information on different types of kriging can be found in Cressie (1991).

4.4 Copula-based Interpolation

This section provides information on how copulas can be used in application to geostatistical interpolation. Copulas are widely used in modern statistics, especially finance, to describe multivariate dependence structure without information about marginal distributions. However, there are only several papers that suggest to use copulas for spatial interpolation. Copulas are extremely helpful because they make possible to model asymmetric dependence and dependence between the quantiles of random variables, in particular tail dependence. Copula is defined as a distribution function on the unit cube $C : [0,1]^p \rightarrow [0,1]$, such that $\forall u_j \in [0,1]^p, j \in (1,\ldots,p)$:

1. $u_j = 0 \Rightarrow C(u_1,\ldots,u_p) = 0$,
2. $C(1,\ldots,1,u_j,1,\ldots,1) = u_j$,
3. $\forall v \in [0,1]^p,\ v_j \leq u_j$

$$\sum_{i_1=1}^{2} \cdots \sum_{i_p=1}^{2} (-1)^{i_1+\cdots i_p} C(g_{1i_1},\ldots,g_{pi_p}) \geq 0$$

for $g_{j1} = v_j$ and $g_{j2} = u_j$.

One of the most important results in the theory of copulas is the Sklars's theorem, which tells how multidimensional distribution, margins and copula are related. For more information about copula and their application in finance we refer to Härdle et al. (2009).

THEOREM 4.1 *Let H be a p-dimensional distribution function with margins $F_1,\ldots,F_p$. Then there exists an p-dimensional copula C such that for all $x = (x_1,\ldots,x_p) \in R^p$*

$$H(x_1,\ldots,x_p) = C\{F_1(x_1),\ldots,F_p(x_p)\}.$$

If $F_1,\ldots,F_p$ are all continuous, then C is unique, conversely, it holds:

$$C(u_1,\ldots,u_p) = H\{F_1^{-1}(u_1),\ldots,F_p^{-1}(u_p)\}.$$

Some efforts of using copula in geostatistics have been already done. Bardossy (2011) suggests to incorporate copulae into the spatial interpolation framework in the following way. Copula C becomes a function of the distance h and does not depend on location in the case of stationarity. Thus, the dependence of any two locations separated by the vector $\|h\|$ is described by

$$\mathrm{P}\{Z(x_i) \leq z_i, Z(x_j) \leq z_j\} = C_h\{F_Z(z_i), F_Z(z_j)\}, \tag{4.10}$$

where F_Z is the univariate distribution of the random process and usually is assumed to be the same for each location x. It is worth to note, that application of the Gaussian copula (see Härdle et al. (2009) for definition) is equivalent to constructing a variogram with the specific covariance structure. Due to this reason kriging can be seen as a special case of the copula interpolation. Bardossy (2011) uses Gaussian and χ^2 copulas. This study takes use of another types of copula, i.e. t-copula, Gaussian and Frank copula. Definitions for all copulas are given in Nelsen (2006).

Gräler et al. (2010) argue that distance has a strong influence on the strength of dependence and introduce concept of bivariate spatial copula. It is assumed that dependence structure is identical for all neighbors, but might change with distance. To construct such a copula it is necessary to build k spacial distance lag classes and estimate bivariate copula density $c_{j,\tau(h)}(u,v)$ for each lag class $[0,l_1),[l_1,l_2),\ldots,[l_{k-1},l_k)$. The function $\tau(h)$ maps distances to a set of estimated parameters. The density of bivariate spatial copula is given by

$$c_h(u,v) = \begin{cases} c_{1,\tau(h)(u,v)} & \text{, if } 0 \leq h < l_1 \\ (1-\lambda_2)c_{1,\tau(h)}(u,v) + \lambda_2 c_{2,\tau(h)}(u,v) & \text{, if } l_1 \leq h < l_2 \\ \vdots & \vdots \\ (1-\lambda_k)c_{k-1,\tau(h)}(u,v) + \lambda_k 1 & \text{, if } l_{k-1} \leq h < l_k \\ 1 & \text{, if } l_k \leq h \end{cases}$$

Herein $\lambda_j = \frac{h-\lambda_{j-1}}{l_j - l_{j-1}}$. Thus, the copula for any new h is evaluated as a weighted sum of copulae for two nearest distances.

To be able to estimate the chosen copula model, it is necessary to transform distribution of $Z(x)$ to the univariate distribution. This can be done using rank transformation $u = \text{rank}\{Z(x)\}/(n+1)$, where n is the sample size. Another way to get uniform marginal distribution is to apply distributional transformation $u = F\{Z(x)\}$, details of such kind of transformations are given in Rüschendorf (2009). Thus, any continious margin can be transformed to be uniformly distributed, but in the case of distributional transform distribution of $Z(x)$ should be known or estimated.

Pebesma et al. (2011) use normal, Student's, lognormal, logistic and generalized extreme value (GEV) distributions as margins. China temperature data have skewed heavy-tailed distribution which is the reason to define margins as the GEV distribution. We refer to Franke et al. (2011) for the precise definition:

$$G_\gamma(x) = \begin{cases} \exp\{-(1+\gamma x)^{-1/\gamma}\} & \text{, if } 1+\gamma x > 0 \text{ and } \gamma \neq 0 \\ \exp\{-\exp(-x)\} & \text{, } x \in R \text{ and } \gamma = 0 \end{cases}$$

The common way to estimate the copula and GEV parameters is maximum likelihood, which is described in Nelsen (2006). Further on, the copula parameter vector will be denoted with $\widehat{\Theta}$.

The final step of the copula-based spatial modeling is interpolation itself. The most popular method for copula-based interpolation is plug-in estimator. The basic idea of the method is to employ the Bayes formula and to get the conditional copula, which is often called h - function, the definition is given in Aas et al. (2006). In the spatial interpolation context, it is given by:

$$c_h\{u(x_0)|\widehat{\Theta}, u(x_1), \ldots, u(x_n)\} = \frac{c_h\{u(x_0), u(x_1), \ldots, u(x_n)|\widehat{\Theta}\}}{c_h\{u(x_1), \ldots, u(x_n)|\widehat{\Theta}\}}, \tag{4.11}$$

where $u(x_i)$ is rank or distributional transformation of the $Z(x_i)$. The predictive density of the $u(x)$ is defined on the unit interval. The density of $Z(x_i)$ should be calculated using the Jacobian transformation. To get back from ranks to the original scale the quantile function F_Z^{-1} is used. The corresponding Jacobian determinant is exactly the density f_z. Hence,

$$\begin{aligned}
\mathrm{E}\{Z(x_0)|\widehat{\Theta}, Z(x_0), \ldots, Z(x_n)\} &= \int_{-\infty}^{+\infty} z(x_0) c_h[F_z\{Z(x_0)\}|\widehat{\Theta}, Z(x_0), \ldots, Z(x_n)]\,\mathrm{d}z(x_0) \\
&= \int_0^1 F_z^{-1}\{u(x_0)\} c_h\{u(x_0)|\widehat{\Theta}, Z(x_0), \ldots, Z(x_n)\}\,\mathrm{d}u(x_0).
\end{aligned}$$

Analogously the prediction's variance can be calculated. For details we refer to Kazianka and Pilz (2010). Sometimes instead of the mean-based prediction is used the median, which is optimal prediction under absolute value loss function.

This research employs the idea used by Patton (2004) to model asymmetric dependence for asset allocation and suggests the extension of the bivariate spatial copula. It is proposed to construct hierarchical copula model, i.e. model the parameter of the copula τ as a function of some geographical characteristics. Another part of extension of the model deals with the model for marginal distributions. In the application to the temperature data it is not possible to use the same margins for the temperature in all locations. The parameters of the GEV distribution are modeled as function of latitude, longitude and elevation, making the model flexible and suitable for spatial temperature interpolation. If we take into account that the temperature is estimated separately for each time moment t, the final model is formulated as follows:

$$\widehat{Z}_t(x_0) = \int_0^1 F^{-1}_{\widehat{\mu}(x_0),\widehat{\sigma}(x_0),\widehat{\xi}(x_0)}\{u(x_0)\} c_{\widehat{\tau}}\{u(x_0)|Z_t(x_k)\}\,\mathrm{d}u(x_0),$$

where

$$\tau = f(h, \alpha, \log\{\Delta(\mathrm{El})\}) + \varepsilon_\tau,$$

$$\mu(x_i) = f(\mathrm{Lat}(x_i), \mathrm{Lon}(x_i), \log\{\mathrm{El}(x_i)\}) + \varepsilon_\mu(x_i),$$

$$\sigma(x_i) = f(\text{Lat}(x_i), \text{Lon}(x_i), \log\{\text{El}(x_i)\}) + \varepsilon_\sigma(x_i),$$

$$\xi(x_i) = f(\text{Lat}(x_i), \text{Lon}(x_i), \log\{\text{El}(x_i)\}) + \varepsilon_\xi(x_i).$$

4.5 IDW-GEV Interpolation

The mentioned above copula-based interpolation approach is supposed to capture extreme temperatures. However, it is computationally intensive. Moreover, it is not clear which part of the model, copula or GEV, is the main contribution to the reduction of error. As alternative to complicated copula modeling the current study suggests to use IDW to get the $\widehat{u}(x_0)$ in unknown location and estimate the temperature using the quantile function of the fitted GEV distribution. To summarize, for each time t the proposed model is:

$$\widehat{Z}_t(x_0) = F^{-1}_{\widehat{\mu}(x_0),\widehat{\sigma}(x_0),\widehat{\xi}(x_0)}\{\widehat{u}_t(x_0)\},$$

where

$$\widehat{u}_t(x_0) = \frac{\sum\limits_{j:\|x_j - x_0\| \leqslant h} w(x_j) u_t(x_j)}{\sum\limits_{j:\|x_j - x_0\| \leqslant h} w(x_j)}. \tag{4.12}$$

$w(x_j) = 1/\|x_j - x_0\|^p$. Remind that $u(x_i)$ is the rank transform of $Z(x_i)$. Parameters of the GEV are taken the same as in 4.4.

5 Results

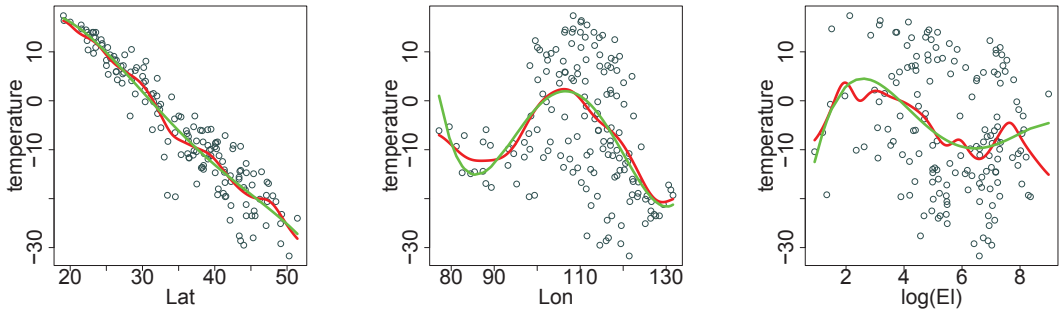

Figure 5.1: Nonparametric and multiple linear regression of temperature on latitude, longitude and logarithm of elevation at $t = 200$.

Regr

A considerable amount of methods were described in Chapter 4. This chapter presents the empirical comparison of all proposed models. They are compared by calculating MAE (see 3.5) averaged over all years for each single station ($i = 1, \ldots, 159$) and averaged over all given stations for each day of the year ($d = 1, \ldots, 365$). MAE is calculated from out-of-sample prediction. For this purpose crossvalidation methodology which is described in Chapter 3 is used.

The first method applied is regression smoothing (see 4.1). Latitude, longitude and logarithm of elevation are included in the model. Figure 5.1 shows scatter plot of coordinates and observed temperature together with the Nadaraya Watson regression and the chosen multiple linear regression model for one concrete moment in time. For the Nadaraya Watson regression Gaussian kernel and plug-in bandwidth estimator are used. For further detail on nonparametric regression we refer to Härdle et al. (2004). This simplified approach was chosen because the main goal of performing nonparametric regression is to select the order of multiple linear regression. The final multiple linear regression model is

$$Z_t(x_i) = \sum_{j=0}^{2} a_{t,j}\text{Lat} + \sum_{j=1}^{4} b_{t,j}\text{Lon} + \sum_{j=1}^{3} c_{t,j}\log(\text{El}) + \varepsilon_t(x_i); t = 1, \ldots, T; i = 1, \ldots, 159.$$

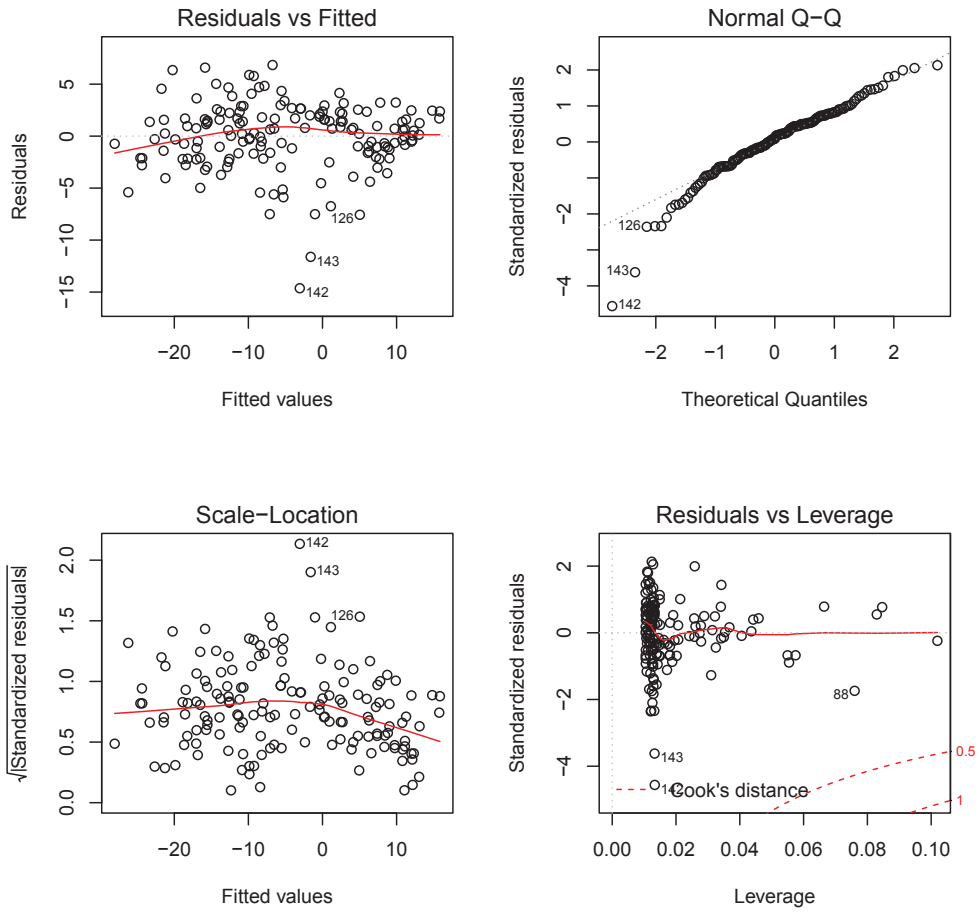

Figure 5.2: Diagnostic plots for multiple linear regression at $t = 200$.

RegrFit

The diagnostic plots for the regression model 5 are shown in the figure 5.2. It can be concluded that the regression fits the data quite well. However, some outliers are observed. They are not excluded from the dataset because extreme temperatures are of great importance in the current research.

It is worth to mention that the goodness of fit and coefficients vary significantly across the time. The R^2 over months are given in the table 5.1. It can be concluded that the regression gives the best fit during the winter months. However, further investigation of the error will show that good fit of the multiple linear regression model does not guarantee small interpolation error. MAE for each station is seen in the figure 5.3. The most obvious finding to emerge from this study is that the linear regression model gives best fit during the summer months (see figure 5.24) and is not extremely sensitive to the location of the station. Thus, latitude,

longitude and logarithm of elevation are reliable predictors of the temperature in unknown location.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
$\overline{R}^2$	0.92	0.91	0.88	0.82	0.75	0.69	0.72	0.77	0.83	0.87	0.92	0.93
SD_{R^2}	0.03	0.03	0.05	0.07	0.09	0.08	0.06	0.06	0.06	0.05	0.03	0.02

Table 5.1: R^2 for the multiple linear regression averaged by month.

Regr

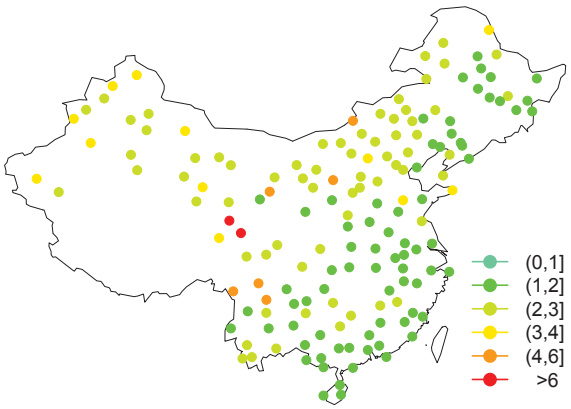

Figure 5.3: MAE for regression model.

MaeRegr

The next method is inverse distance weighting presented in 4.1. In contrast to usual approach, this study suggests to choose p and h locally, such that MAE is minimized for each station:

$$h_i = \arg \min_{h \in [Q_{0.05}, Q_1]} \Sigma_{t=1}^{T} |Z_t(x_i) - \widehat{Z}_t(x_i)|,$$

$$p_i = \arg \min_{p \in [0.5, 20]} \Sigma_{t=1}^{T} |Z_t(x_i) - \widehat{Z}_t(x_i)|.$$

This research uses distance-based nearest neighbor method which results in more precise prediction. p is chosen from the sequence $[0.5, 20]$ with the step 0.5, h takes values from the quantiles of the distribution of all separating distances given in the table 5.2. Figure 5.4 is evidence to the fact that the choice of these two parameters has influence on the interpolation error. The figure 5.5 shows the optimal p and h for each station. If these values are used,

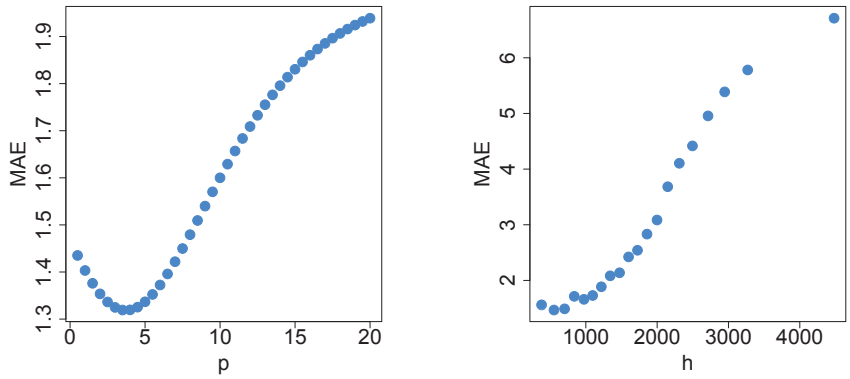

Figure 5.4: Optimal p (left) and h (right) for station $i = 26$ and $t = 200$.

OptPar26

	Min	Q1	Median	Mean	Q3	Max	SD
h	30.88	976.05	1600.67	1683.13	2312.26	4480.88	887.42

Table 5.2: Numerical summary for distances between the stations.

DistSta

the method is able to interpolate the temperature with minimal error. However, in practice the optimal parameters are unknown for unknown location. The problem has two possible solutions. The first is to choose parameters which are optimal for all stations. The second is to take the distance weighted mean of parameters of closely located stations. The investigation of this question shows that these two approaches give similar errors. Intuitively it is clear from the plot 5.5. There is no spatial patter in p and h. Two stations which are situated very close to each other may have extremely different parameters. $p = 3$ and $h = 553$ are parameters which minimize the MAE over all stations. Figure 5.6 illustrates the MAE for each station. The results of this study indicate that IDW performs well in the coastal area where temperature fluctuations are not wide and stations are located closer to each other. On the contrary, MAE values in some areas of the continental part reaches the value of $8\,^{\circ}\mathrm{C}$.

The next method applied to the China temperature data set is ordinary kriging. First, the 4 directional variograms for $\alpha = \{0, \pi/4, \pi/2, 3\pi/4\}$ are constructed and shown in the figure 5.7. This preliminary analysis suggests to choose two main directions, 0 and $\pi/2$. Variograms for these direction with the fitted Gaussian model are shown in the picture 5.8. This figure indicates that the main direction is $\alpha = \pi/2$ and the dependence in direction from the north to the south is much stronger than in direction from the east to the west. This conclusion is drawn from the observed sill value (see 4.3 for definition). By analogy to the IDW, when the

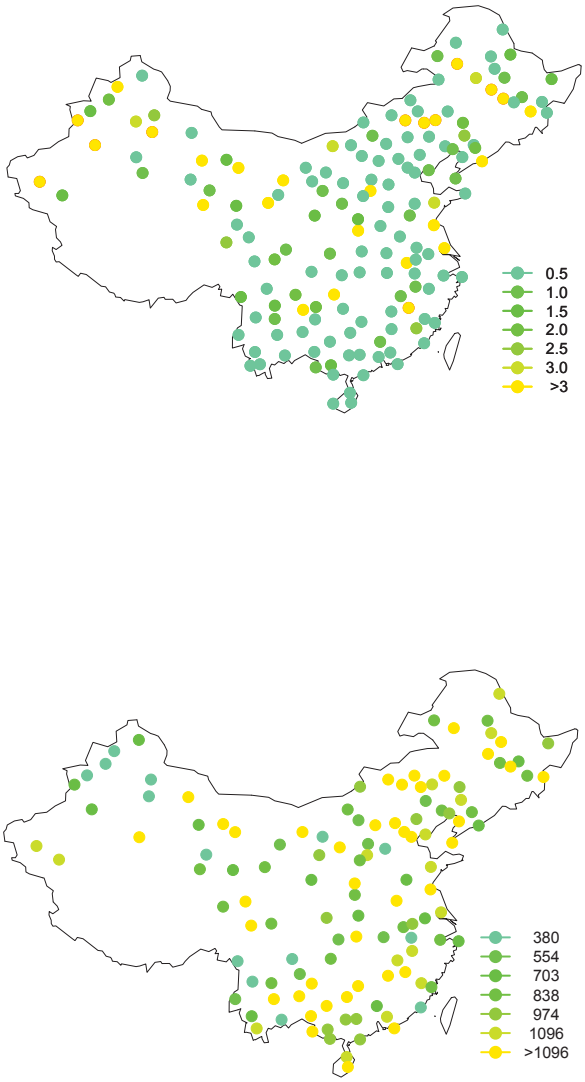

Figure 5.5: Optimal p (top) and h (bottom) for each station.

OptPar

model is estimated, it is necessary to choose the maximal distance h in 4.7. This means that only observations within a distance h from the unknown location are used for prediction. h is chosen as in IDW from quantiles of the distribution of all separating distances. MAE is strongly influenced by h. This finding coincides with the results of IDW interpolation. Thus, the smaller h are preferable in order to reduce the interpolation error. However, the influence

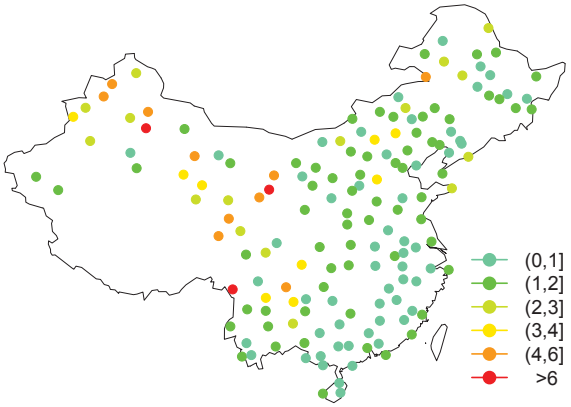

Figure 5.6: MAE for IDW model.

MaeIdw

of h in kriging is not such strong as in IDW. It is observed that MAE varies for small h values and is almost constant for $h > 1000$. For the constructing of the final model $h = 1096$ was chosen. MAE for all stations is given in the picture 5.9. The research has shown that the universal kriging model does not improve the spatial prediction. Lat, Lon and $\log(\text{El})$ were added to the model as covariates for the surface trend function. Contrary to expectations, this study did not find a significant improvement of the model. There is possible explanations to this result. The information about geographical coordinates is already captured by the distance. This conclusion is supported by the closer look at the variogram 5.8. It is seen that the nugget is almost zero, which means that spatial trend and dependence can be captured by the ordinary kriging.

The next part of this chapter contains the discussion of copula-based interpolation technique. It is worth to emphasize that there are two possible applications of copulas in the context of spatial data. The first one is determining the dependence among concrete spatial locations. For this purpose Xu et al. (2010) propose to use residuals of the temperature after removing time trend. This approach is useful for diversification problem, but faces difficulties applied to the spatial interpolation problem. In such a situation the output of the model is a set of residuals in unknown locations. Cao (2012) argues that it is possible to make interpolation of the residuals and then predict the value of the process using e.g. AR process. However, the time series models are different for each location in space. Consequently, they are unknown for the new locations in space. This leads to the additional interpolation of parameters of AR process and can lead to huge errors. That is the reason for interpolating temperature data, not the residuals.

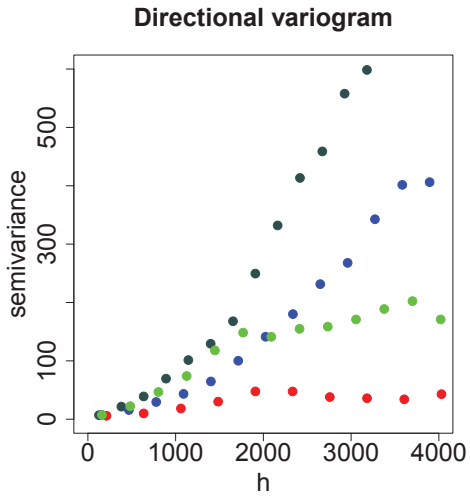

Figure 5.7: Directional variogram for $\alpha \in \{0, \pi/4, \pi/2, 3 \cdot \pi/4\}$.

DirVar

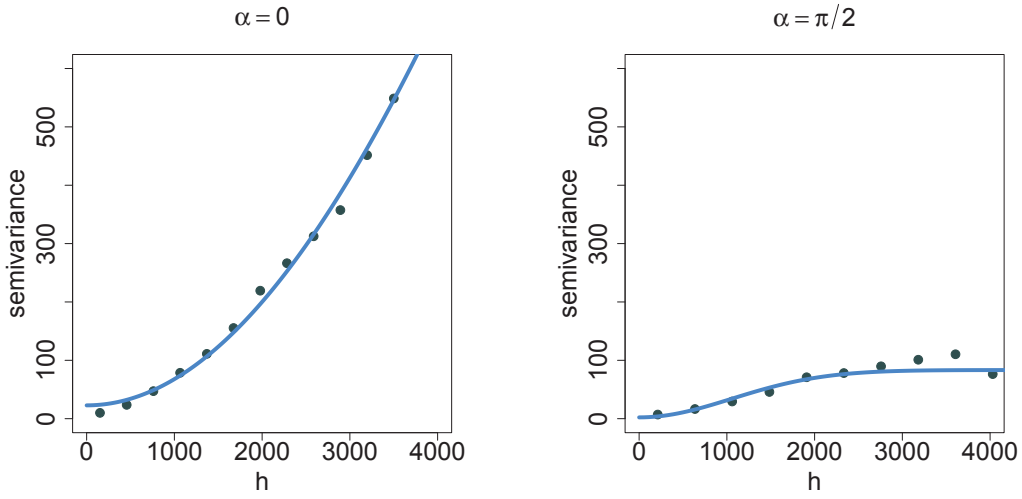

Figure 5.8: Directional variogram for $\alpha \in \{0, \pi/2\}$.

DirVar

In was mentioned earlier that Bardossy (2011) proposes to use the bivariate copula and generalized extreme value distribution with fixed parameters. In contrast to all previous models this study proposes to model copula parameter as a function of separating distance, angle and elevation. GEV is also assumed to be flexible, i.e. its parameters are modeled as functions of geographical characteristics. GEV distribution is chosen to capture extremely law and high temperatures. In addition, it can describe changing skewness of the distribution which depends on the season.

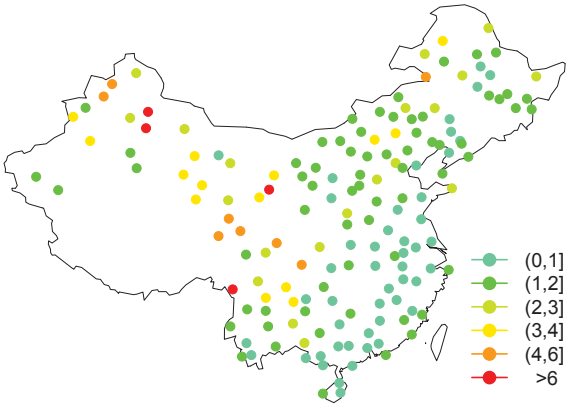

Figure 5.9: MAE for kriging model.

MaeKrig

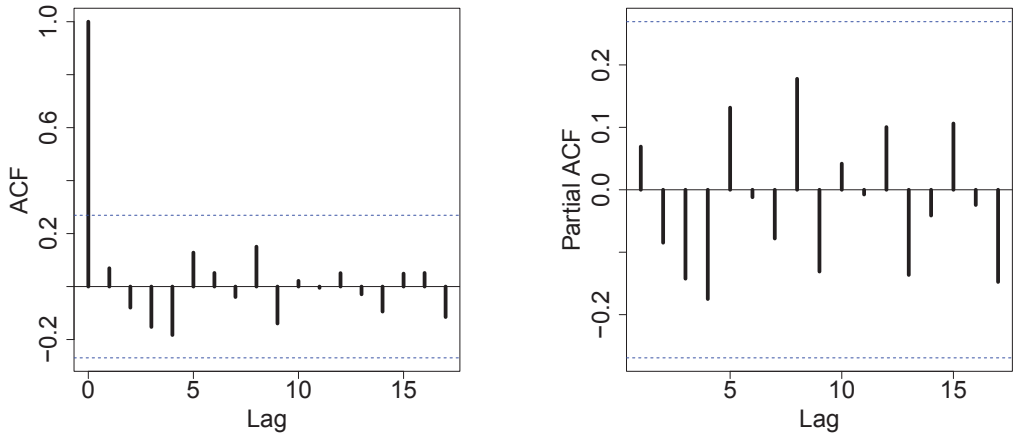

Figure 5.10: ACF (left) and PACF (right) of temperature for $i = 26$ and $d = 18$th of July. ADF test p-value < 0.01, Ljung-Box test p-value $= 0.61$.

AcfPacf

However, before apply copula and GEV are applied it is necessary to insure that there is no serial dependence in the data. It is not possible to apply this methodology directly to the daily temperature data. To avoid this problem the following trick is proposed. For each station all observation are grouped by the day of the year, i.e. all 1st of January, all 2nd of January and so on. Further on, the day of the year is denoted as $d = t - \lfloor t/365 \rfloor \cdot 365 = t \bmod 365$. Thus, 365 groups for each station are obtained . Each group contains 53 observations. The hypothesis to

be tested is that the observations within each group are not serially correlated. The ACF and PACF for temperature and squared temperature in the station $i = 26$ at day $d = 18$th of July (figures 5.10 and 5.11) suggest not to reject the mentioned above hypothesis. However, more precise tests are done. p-value for ADF test, H_0: there is unit root in the given time-series sample, test is smaller than 0.01. p-value of Ljung-Box test, H_0 : there is no serial correlation in the given time series sample, is 0.61. Therefore the hypothesis can not be rejected. The same is true for the majority of the stations and days of the year. Consequently, the first part of the copula-based interpolation, i.e. modeling of the marginal distributions, can be done.

The algorithm for copula-based interpolation can be formulated as follows:

- Estimate marginals
  - Estimate GEV parameters for each station and each day of the year
  - Model dependence of GEV parameters from geographical coordinates (use multiple linear regression)

- Estimate copula
  - Choose bivariate copula
  - Estimate copula parameter for each pair of stations
  - Model copula parameter as function of separating distance h and angle α

- Calculate $u_t(x_i) = [\text{rank}\{Z_d(x_i)\}/54]_{(t \text{ div } 365)}$ as rank transformation of the temperature in the station i at day d

- Estimate $\widehat{Z}_t(x_0)$ from 5

First, the parameters of the GEV distribution are estimated: location, shape and scale parameter. This is done for each station-day group. Remind, that each group has 53 observations. Figure 5.12 demonstrates on the example of station $i = 26$ and day $d = 18$th of July that the chosen distribution type fits the data well. The same is true for all station-day groups. The next step is to model parameters of the GEV distribution as linear function of the geographical coordinates. This is done for all $d = 1, \ldots, 365$. It was not possible to fit joint model for all days. The reason for this could be the different effect of time for all the stations. More complicated models which included such covariates as $d \cdot \text{Lat}$, $d \cdot \text{Lon}$ etc. failed as well. Due to strong influence of time variable all other parameters were estimated biased. As in the regression analysis the Nadaraya Watson regression was used as a tool for selecting the right order of the multiple linear regression. Nonparametric estimate and multiple linear regression fit can be seen in the pictures 5.13, 5.14 and 5.15.

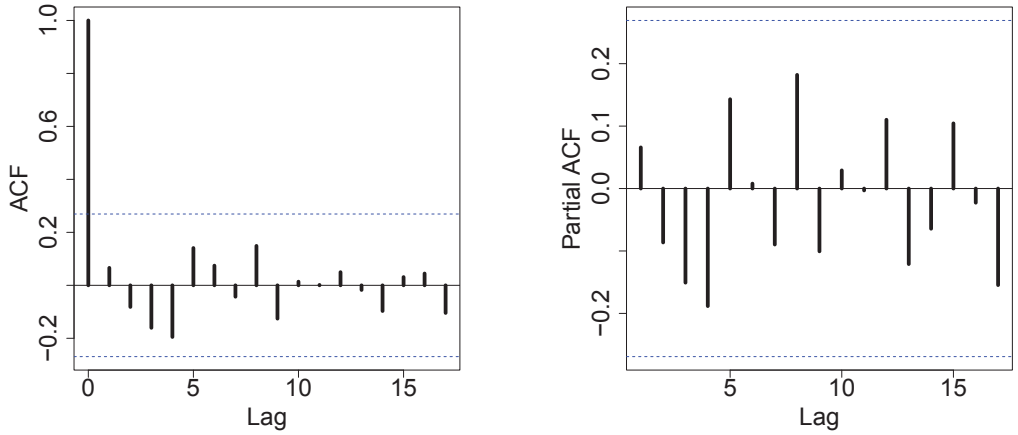

Figure 5.11: ACF (left) and PACF (right) of squared temperature for $i = 26$ and $d = 18$th of July.

AcfPacfSq

The chosen model for the parameters of the GEV distribution are given below:

$$\mu_d(x_i) = \sum_{j=0}^{2} a_{\mu,d,j}\text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\mu,d,j}\text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\mu,d,j}\log\{\text{El}(x_i)\}^j + \varepsilon_d(x_i)$$

$$\sigma_d(x_i) = \sum_{j=0}^{3} a_{\sigma,d,j}\text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\sigma,d,j}\text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\sigma,d,j}\log\{\text{El}(x_i)\}^j + \varepsilon_d(x_i)$$

$$\xi_d(x_i) = \sum_{j=0}^{4} a_{\xi,d,j}\text{Lat}(x_i)^j + \sum_{j=1}^{3} b_{\xi,d,j}\text{Lon}(x_i)^j + \sum_{j=1}^{3} c_{\xi,d,j}\log\{\text{El}(x_i)\}^j + \varepsilon_d(x_i)$$

The third step in the copula-based interpolation is copula modeling. First, it is necessary to choose the copula type and estimate copula parameter. Then, copula parameter should be modeled as a function of separating distance, angle and logarithm of the difference in elevation. To choose the copula type contour plots for 3 closely located stations were constructed (see figure 5.16). This visual analysis suggests to use Frank copula. Although, elliptical copulas, Gaussian and t, may already give the satisfactory fit. More deep investigation of this problem shows that there is very small difference in prediction using these 3 different copulae. However, model fails when the independence copula is used. Figure 5.17 shows the comparison of prediction given by 4 different copula models for the station $i = 139$ and year 1980. Gaussian copula is chosen for the final model due to the simplicity of estimation. The parameter is estimated for all pairs of stations at each $d = 1, \ldots, 365$. Only stations which are situated closer to each other than 1000 km are taken into consideration. The first reason for excluding other pairs is rapid growing of variation of the estimate for the bigger distances.

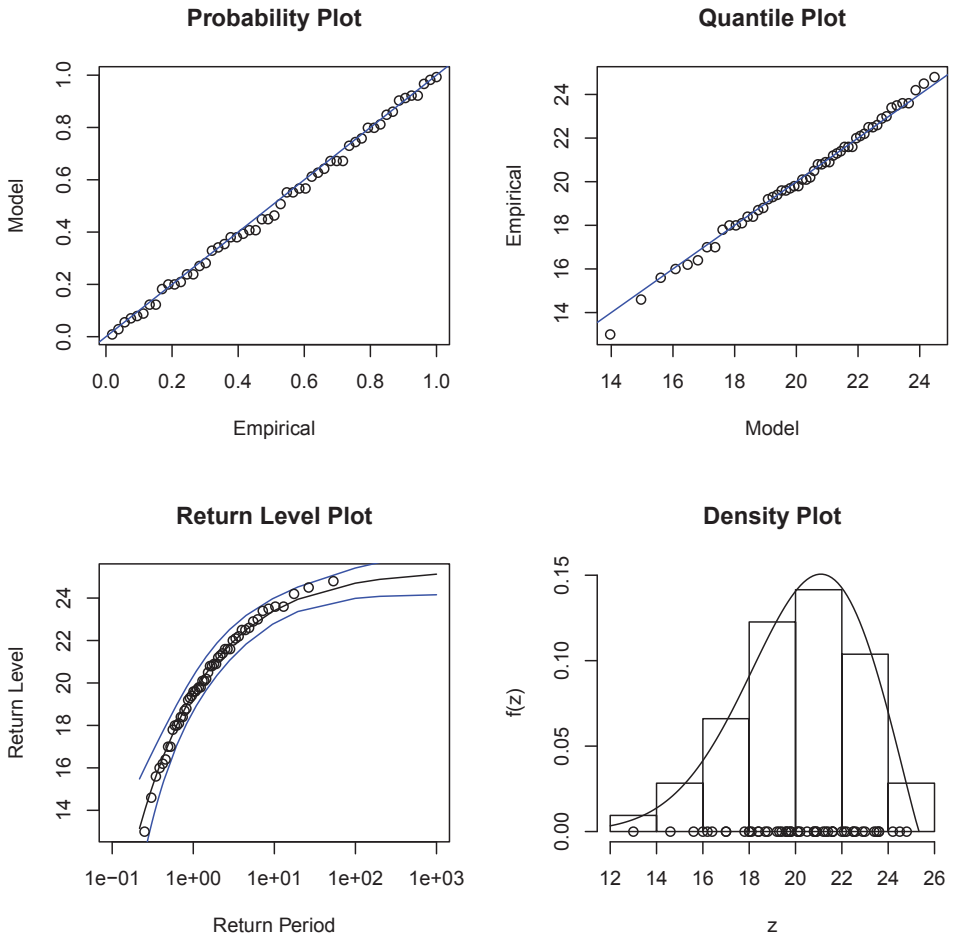

Figure 5.12: Goodness of fit for GEV distribution ($i = 26$ and $d = 18$th of July).
GevFit

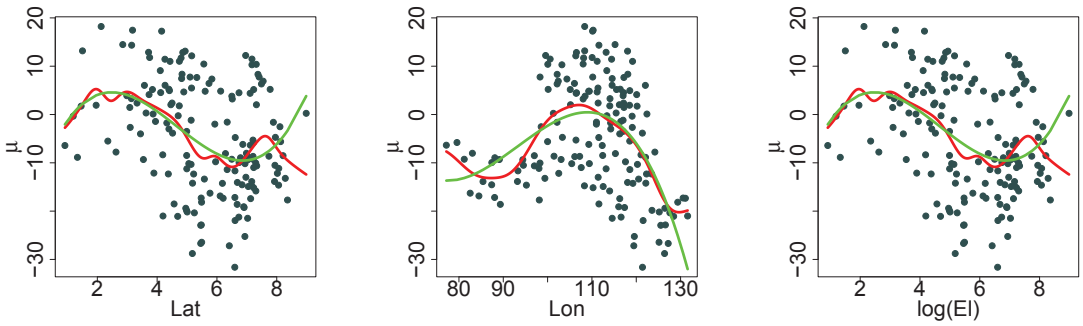

Figure 5.13: μ_{200} as nonparametric and multiple linear regression of Lat, Lon and $\log(\text{El})$.
GevMu

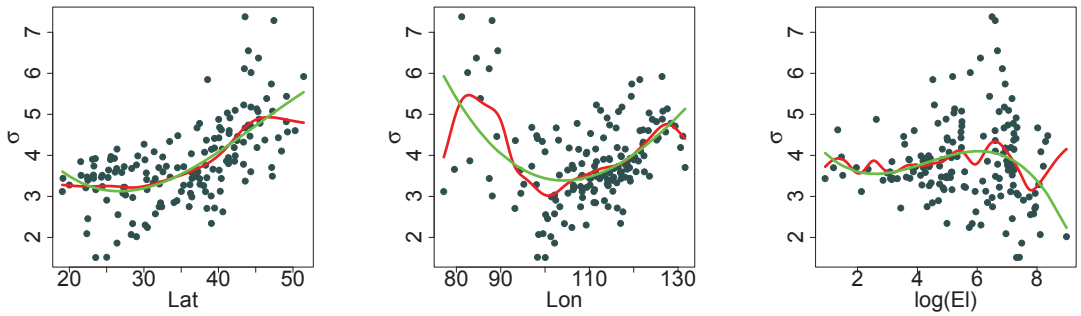

Figure 5.14: σ_{200} as nonparametric and multiple linear regression of Lat, Lon and log(El).
GevSigma

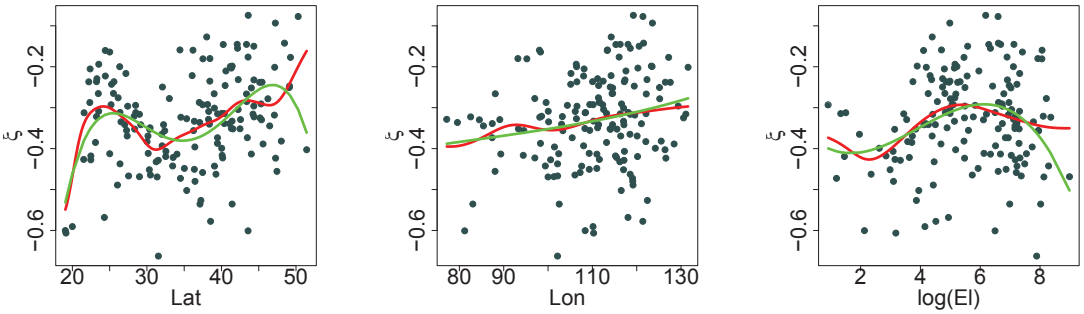

Figure 5.15: ξ_{200} as nonparametric and multiple linear regression of Lat, Lon and log(El).
GevXi

This fact makes the model of the parameter more difficult. Besides this, IDW and kriging analysis showed that more distant locations should not be taken into account. Remind that 53 pairs of observations are available to estimate each bivariate copula. Estimated parameters with fitted Nadaraya Watson and multiple linear regression are seen in the figure 5.18. This figure indicates that the copula parameter is mainly influenced by the distance between the stations. The final model for copula parameters is given by:

$$\rho_d = \sum_{j=0}^{2} a_{\rho,d,j} h^j + \sum_{j=1}^{3} b_{\rho,d,j} \alpha^j + \sum_{j=1}^{3} c_{\rho,d,j} \log\{\Delta(\mathrm{El})\}^j + \varepsilon_d.$$

When all the parameters are estimated it is possible to obtain prediction at unknown location as:

$$\widehat{Z}_t(x_0) = \int_0^1 F^{-1}_{\widehat{\mu}_d(x_0),\widehat{\sigma}_d(x_0),\widehat{\xi}_d(x_0)}\{u(x_0)\} c_{\widehat{\rho}_d}\{u(x_0)|Z_t(x_k)\}\,\mathrm{d}u(x_0).$$

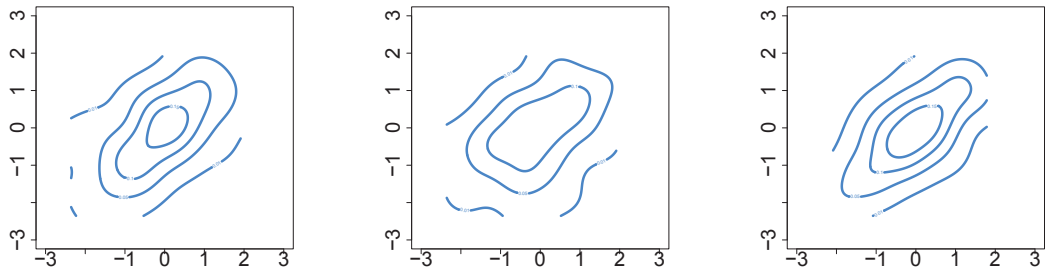

Figure 5.16: Empirical contour plots for 3 pairs of stations.

MetCon

Usual crossvalidation procedure is performed for this method. Out-of-sample MAE is displayed on the figure 5.19. If this figure is compared to 5.6 it can be concluded that copula-based interpolation reduces the error in the continental part of the region and is able to capture extreme temperatures. In contrast, this method is outperformed by the IDW in the coastal areas.

The aim of the further research is to keep flexibility of the IDW model and employ the ability of GEV to model extremely high or law temperatures. Let us take a closer look at the copula-based interpolation model. There are several sources of error in this model. The main error comes from modeling the copula parameter. Another major source of uncertainty is a result of modeling GEV distribution. It is of crucial importance to understand which step of the copula-based interpolation algorithm results in a higher prediction error. For this purpose two kinds of errors are calculated. First of all, MAE for all stations is calculated using estimated copula parameter and fitted values of GEV distribution's parameters. Secondly, MAE is determined by taking estimated (not multiple linear regression fitted) GEV distribution's parameters and evaluating copula parameter with the copula model. Results are shown on the figure 5.20. It is obvious that copula part of the model is the main source of error. The upper part of the figure leads to the following conclusion - if copula parameter and consequently $u_t(x_i)$ are fitted in a proper way, it is possible to reduce error in the mountain area and keep it small in the coastal area. Figure 5.21 shows the pattern of $u_{200}(x_i)$, which form some spatial clusters. There are areas where the $u_{200}(x_i)$ are closer to 0 and areas where $u_{200}(x_i)$ are almost 1. It follows that $u_{200}(x_i)$ at unknown location can be estimated as a weighted average of the nearest neighbors. This conclusion gives the main idea for the new method - IDW-GEV interpolation. It proposes to estimate $u_{200}(x_i)$ at unknown location using IDW methodology and than transform it to the temperature employing the quantile function of the

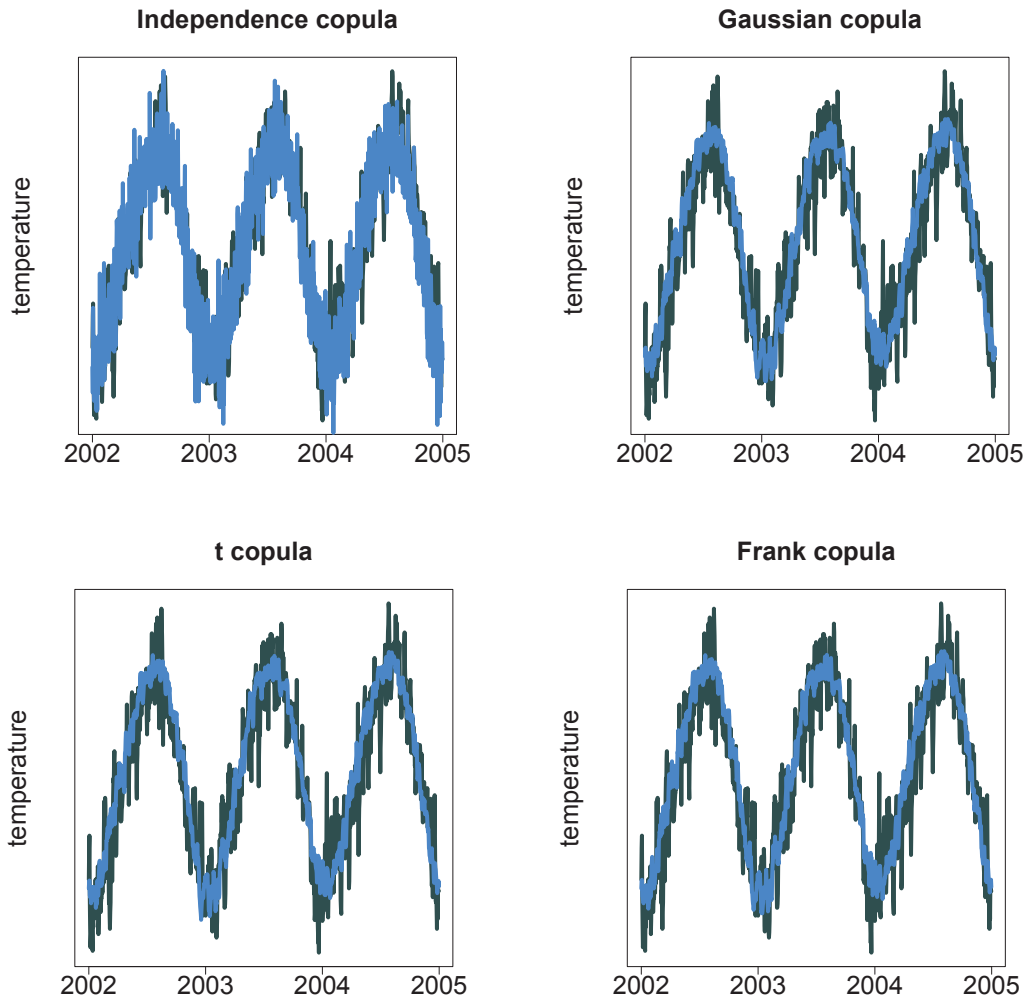

Figure 5.17: Temperature prediction given by different kind of copulas.
CopCom

GEV distribution. This can be formalized as follows:

$$\widehat{Z}_t(x_0) = F^{-1}_{\widehat{\mu}_d(x_0),\widehat{\sigma}_d(x_0),\widehat{\xi}_d(x_0)}\{\widehat{u}_t(x_0)\},$$

where

$$\widehat{u}_t(x_0) = \frac{\sum_{i:\|x_j - x_0\| \leqslant d} w(x_j) u_t(x_j)}{\sum_{i:\|x_j - x_0\| \leqslant d} w(x_j)}$$

and $w(x_j) = 1/\|x_j - x_0\|^p$, $u_t(x_i) = [\text{rank}\{Z_\tau(x_i)\}/54]_{(t \text{ div } 365)}$. $\widehat{\mu}_d(x_i), \widehat{\sigma}_d(x_i), \widehat{\xi}_d(x_i)$ as in the copula interpolation formula. t div $365 = \lfloor t/365 \rfloor$. Optimal p and h are chosen minimizing the MAE over all stations. Thus, $p = 0.5$ and $h = 553$. MAE for the IDW-GEV method

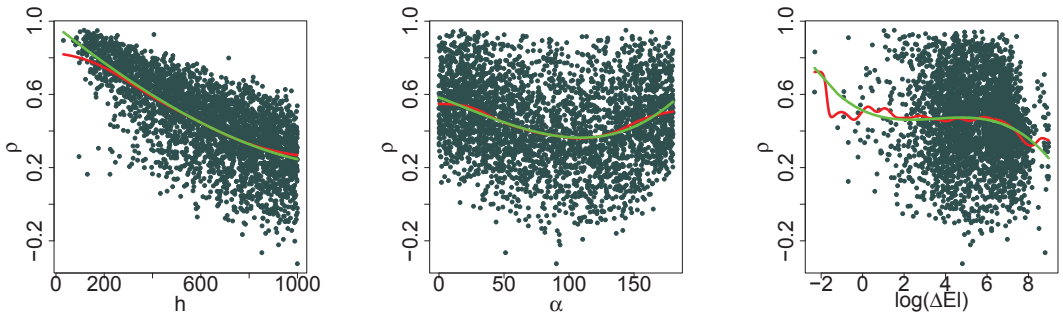

Figure 5.18: Gaussian copula parameter as nonparametric and multiple linear regression on separating h, α and logarithm of elevation difference $\log\{\Delta(\text{El})\}$.

CopRegr

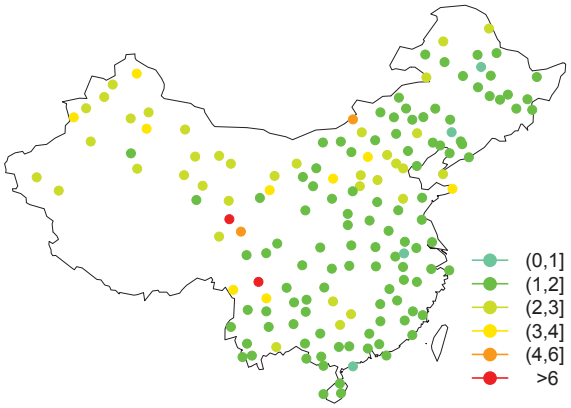

Figure 5.19: MAE for copula-based interpolation model.

MaeCop

for all stations is given in the figure 5.22. As was expected the IDW-GEV method is able to capture the extreme observations in the continental part of China and does not cause in higher MAE in the coastal area. Error is reduced in at least 50 % of the locations. This number varies according to the season and reaches its maximum during the summer period. Surprisingly, the improvement was found to be not location dependent. It is possible to hypothesize that error is reduced in the mountain areas and by coastline. However this should be tested in more detail.

The result of this investigation suggests two competing methods : IDW and IDW-GEV. It depends on the concrete application which one should be preferred. Figure 5.23 demonstrates

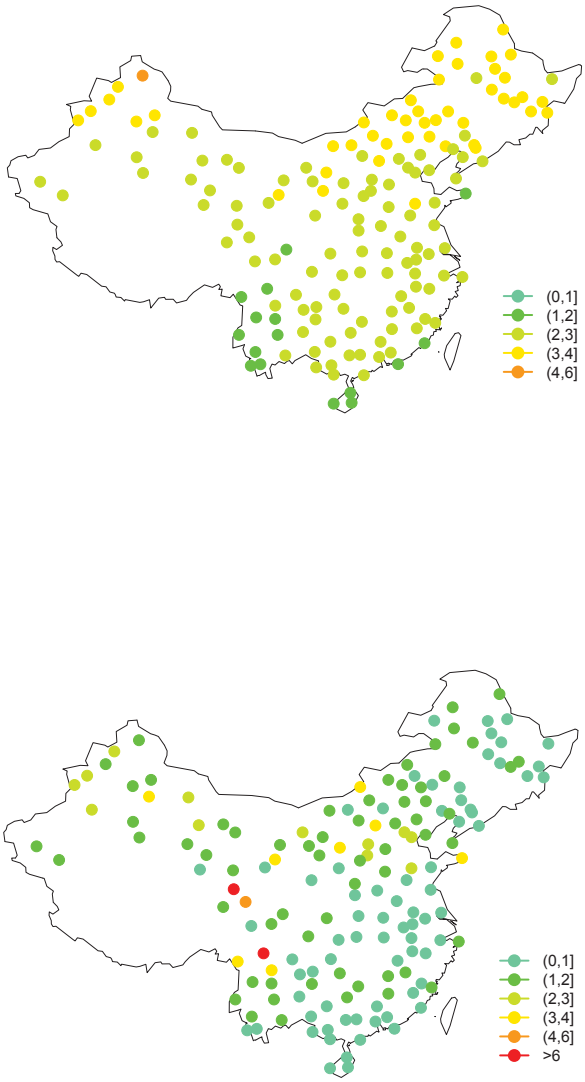

Figure 5.20: MAE for copula-based interpolation model with fitted copula parameter and estimated GEV distribution's parameters (top) and estimated copula parameter and fitted GEV disstribution's parameters (bottom).

EstFit

the performance of all described interpolation techniques for the station number $i = 139$ (Dulan) which is located in the central part of China. This simple visual analysis discovers that IDW is not successful in predicting extreme temperatures and is underestimating both high

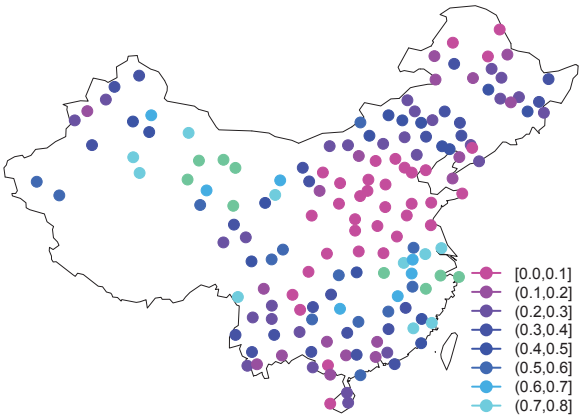

Figure 5.21: $u_t(x_i)$ pattern at $t = 200$.

Upat

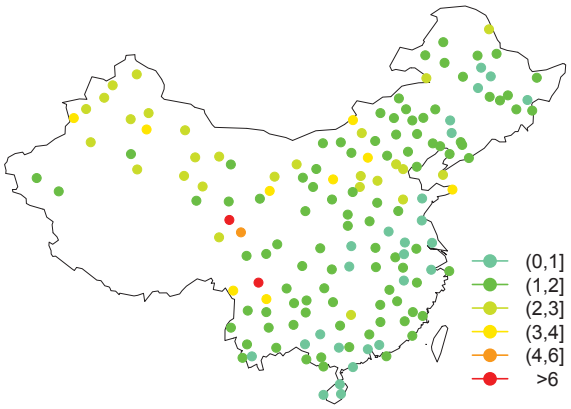

Figure 5.22: MAE for IDW-GEV interpolation model.

MaeIdwgev

and low temperatures. It can be concluded that in cases when not only the mean error is of high importance IDW-GEV methodology outperforms local IDW interpolation. In addition, it is interesting to note that all methods have season dependent error. Figure 5.24 is an illustration of this phenomena. It shows mean MAE over all stations for all days of the year. MAE reaches its maximal value during the winter period and comes down in the summer, which is

quite useful for applications in agriculture.

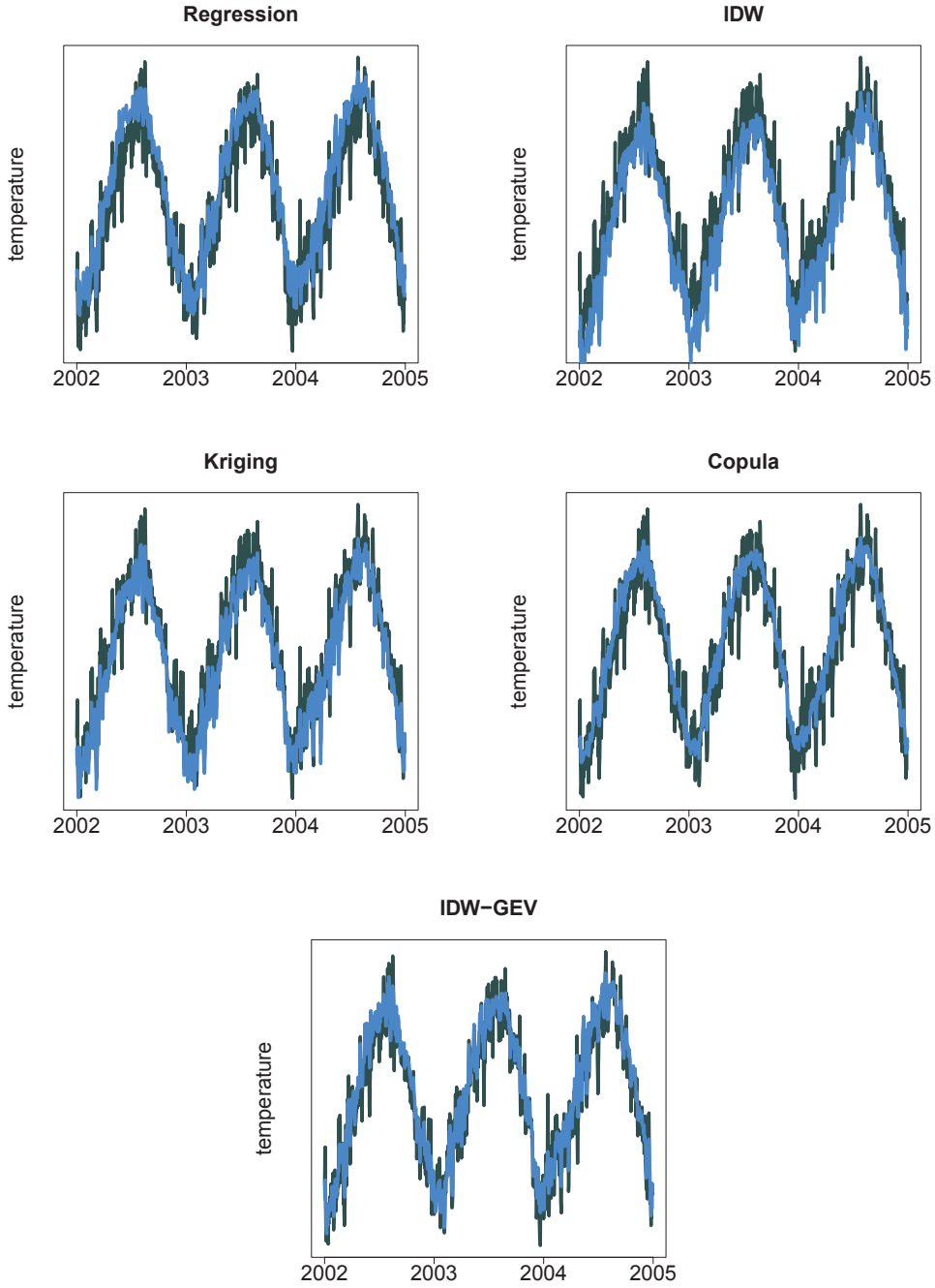

Figure 5.23: Regression, IDW, kriging, copula, IDW-GEV prediction for station $i = 139$.
Ts139

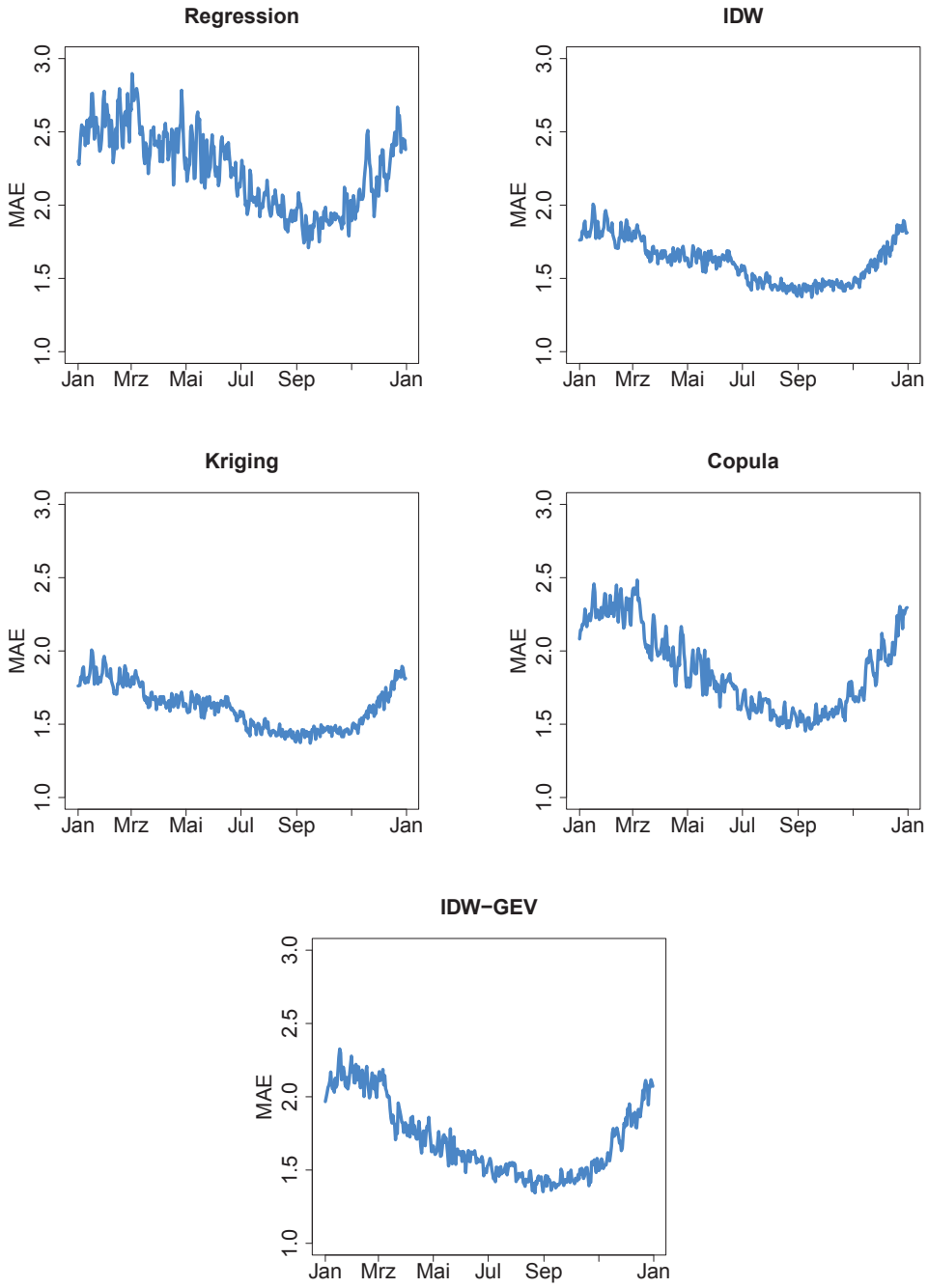

Figure 5.24: MAE for regression, IDW, kriging, copula, IDW-GEV interpolation models averaged by the day of the year.

TsMae

6 Conclusions

This work gives the application of the spatial interpolation techniques to the daily average temperature in China in the time period from 1957 to 2009. Three methods described in the literature were adopted to the peculiarity of the climate in China. The parameters of inverse distance interpolation and ordinary kriging were estimated in order to minimize the interpolation error. The regression model was modified to polynomial regression. The present study extended the bivariate copula spatial model and proposed to model the parameters of the bivariate copula as a continuous function of separating distance, angle and elevation. The model of margins was adopted to the temperature data as well, making its parameters location dependent. One of the interesting findings of this study was the fact that the different types of copulae resulted in almost the same prediction. The simplification of copula-based interpolation, IDW-GEV interpolation, was designed. All the models were compared estimating out-of-sample prediction and computing the absolute average error.

The findings of this research suggested that all mentioned above interpolation techniques gave time and space dependent error, which resulted in the loss of precision during the winter months and in the mountain regions of China. The regression model was found to give the biggest error, kriging model gave better results but was outperformed by other methods.

The innovative copula model was found to be able to capture extreme observations and to reduce the prediction error in the mountain areas, however, resulted in higher error in the coastal line. It was shown that the copula model can be simplified in IDW-GEV model, which is less computing intense.

Local inverse distance interpolation and IDW-GEV emerged as reliable models for temperature interpolation. Inverse distance interpolation performed better than IDW-GEV during the winter period giving maximal average absolute error $2\,^{\circ}\mathrm{C}$. However, it failed in the prediction of extreme temperatures. IDW-GEV gave improvement to the prediction in about 50 % of stations. During the summer and autumn period both models gave similar results with minimal error of $1.3\,^{\circ}\mathrm{C}$ in September. Thus, the model to be preferred depends on the concrete applications. If the minimal average error is of main interest, local inverse distance interpolation should be used, however, if extremes are of great importance, IDW-GEV should be preferred.

Bibliography

K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. Sonderforschungsbereich 386, Paper 487, 2006.

O. Babak and C. Deutsch. Statistical approach to inverse distance interpolation. *Stochastic Environmental Research and Risk Assessment*, 23(5), 2008.

A. Bardossy. Interpolation of groundwater quality parameters with some values below the detection limit. *Hydrology and Earth System Sciences*, 8(3), 2011.

R. Bivand, E. J. Pebesma, and V. Gomez-Rubio. *Applied Spatial Data Analysis with R.* Springer, 2008.

C. Brechmann and U. Schepsmeier. Modeling dependence with C- and D-vine copulas. 2011.

X. Cao. Water level modeling around German bight, 2012.

H. Chai, W. Cheng, C. Zhou, X. Chen, X. Ma, and S. Zhao. Analysis and comparison of spatial interpolation methods for temperature data in Xinjiang Uygur autonomous region, china. *Natural science*, 3(12), 2002.

N. A. C. Cressie. *Statistics for Spatial Data.* John Wiley & Sons, 1991.

P. Diggle and P. Ribeiro. *Model-based Geostatistics.* Springer, 2007a.

P. J. Diggle and P. Ribeiro. geoR : Package for geostatistical data analysis. an illustrative session. 2007b.

J. Franke, W. K. Härdle, and C. Hafner. *Statistics of Finantial Markets.* Springer, 2011.

C. Gaetan and n. X. Guyo. *Spatial Statistics and Modeling.* Springer, 2010.

M. G. Gentona and D. J. Gorsich. Nonparametric variogram and covariogram estimation with Fourier–Bessel matrices. *Computational Statistics & Analysis*, (47), 2002.

M. G. Gentona and D. J. Gorsich. On nonparametric variogram estimation. *Journal of the Korean Statistical Society*, 41(3), 2012.

W. H. Green. *Econometric Analysis.* Springer, 2011.

B. Gräler, H. Kazianka, and G. M. de Espindola. Copulas, a novel approach to model spatial and spatio-temporal dependence. 2010.

W. K. Härdle and L. Simar. *Applied Multivariate Statistical Analysis.* Springer, 2012.

W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models.* Springer, 2004.

W. K. Härdle, N. Hautsch, and L. Overbeck. *Applied Quantitative Finance.* Springer, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

M. Holdaway. Spatial modeling and interpolation of monthly temperature using kriging. *Climate Research*, 6, 1992.

H. Kazianka and J. Pilz. Spatial interpolation using copula-based geostatistical models. *Quantitative Geology and Geostatistics*, 16, 2010.

D. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 21(4), 1951.

H. Lauren, R. Viger, and G. McCABE. Precipitation interpolation in mountainous regions using multiple linear regression. *Hydrology, Water Resources and Ecology in Headwaters*, (248), 2002.

M. Loecher. Plotting on google static maps in R. 2010.

M. Loecher. Spatio-temporal analysis and interpolation of PM10 measurements in Europe. ETC/ACM Technical Paper 2011/10, 2011.

R. B. Nelsen. *An Introduction to Copulas.* Springer, 2006.

A. Patton. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal ofFinancial Econometrics*, 2(1), 2004.

E. Pebesma. The meuse data set : a tutorial fot the gstat r package. 2006.

E. Pebesma, D. Cornford, G. Dubois, G. Heuvelink, D. Hristopoulos, J. Pilz, U. Stohlker, G. Morin, and J. Skoien. INTAMAP: the design and implementation of an interoperable automated interpolation web service. *Computers & Geosciences*, 37(3), 2011.

L. Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. 2009.

M. Sherman. *Spatial Statistics and Spatio-Temporal Data. Covariance Functions and Directional Properties.* John Wiley & Sons, 2011.

W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 53(2), 2008.

L. Waller and C. Gotway. *Applied Spatial Statistics for Public Health Data.* John Wiley & Sons, 2004.

R. Webster and M. Oliver. *Geostatistics for Environmental Science.* John Wiley & Sons, 2007.

W. Xu, O. Okhir, M. Odening, and C. Ji. Systemic weather risk and crop insurance: The case of China. Working paper, SFB 649 Humboldt Universität zu Berlin, 2010.

J. Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 52, 2007.

Declaration of Authorship

I hereby conrm that I have authored this master thesis independently and without use of others than the indicated sources. Where I have consulted the published work of others, in any form (e.g. ideas, equations, gures, text, tables),this is always explicitly attributed.

Berlin, 12.11.2012 Anastasija Tetereva

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 12.11.2012 Anastasija Tetereva